

A squandered opportunity

The sacking of France's science and education minister was unavoidable. His successor would do well to take on board some of Claude Allègre's goals, but also to learn from his mistakes in reforming French and European research.

A French newspaper joked this week that Claude Allègre, France's departing minister for research and education, might prefer to avoid the traditional courtyard ceremony of handing power over to his successor, in case ministry staff leaned out of their windows to applaud his departure. The jab is unkind. But it speaks volumes about Allègre's uncanny ability to make himself unpopular, a trait directly responsible for his downfall. Prime minister Lionel Jospin did not part lightly with his lifelong friend, but Allègre's arrogance and inability to consult or negotiate effectively were the anathema of the modern 'listening' government that Jospin aspires to create. In practice, by provoking defensive hostility in others, Allègre reinforced the very conservatism he rightly set out to overcome. The result has been deadlock on most important fronts. In the end, Allègre so isolated himself from the grass roots of the teaching profession that he no longer had the authority to secure change, and Jospin had no choice but to dispose of him. In the circumstances, it is a relief that the option of retaining Allègre in a narrower role as research minister was not pursued.

Entering office in 1997, Allègre recognized the problems of French research as clearly as anyone, and had some excellent goals. Last year's law on innovation, thrashed out with the then industry minister, Dominique Strauss-Kahn, represents a major step towards removing obsolete restrictions on public scientists' involvement with private companies. Also praiseworthy were Allègre's plans to shift resources towards investigator-driven research, and to encourage greater autonomy among young researchers — who, under the current feudal French laboratory system, are tied to their laboratory head's apron-strings. Funding has improved slightly, some jobs have been created, and all high schools are now connected to the Internet.

Wrong reforms

But so much more needed to be achieved, and, given the underlying support for change in the research community, the blame for failures rests largely with Allègre. Instead of the pragmatic modernization of practices, each tuned to particular situations at the grass roots, the ministry chose high-profile reforms of, for example, whole funding agencies, usually preceded by a blitzkrieg on their staff and practices — from the fat "mammoth" of the education system to the "chloroformed" trade unions. When teachers struggling with bursting classrooms called for more posts, the ministry unfairly accused the profession of excessive absenteeism. Reform, often simplistically modelled on the Anglo-Saxon system, seemed often to become an end in itself. Rhetoric replaced focused objectives and detailed means of achieving them.

The pattern became familiar: bold but poorly thought-out reforms were rolled out with fanfare, researchers took to the boulevards, the ministry backed down, and everyone agreed to unambitious plans that could have been reached easily and quickly — witness this week's reform of the Centre National de la Recherche Scientifique (see page 426). Valuable time and energy have been lost as a result of sterile confrontations between researchers and a ministry that appeared to be unable to listen. A more productive

strategy would have been for the ministry and the research community to identify critical areas where tangible progress could be achieved. As a result, three years after coming to power, the government's overall achievements in research reform could be hidden under a pebble beneath the virtually unbroken palisades of France's research institutions.

Power problems

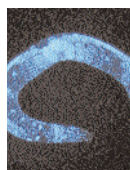
Roger-Gérard Schwartzberg, Allègre's successor as research minister, would do well to study the causes of this failure. Apart from Allègre's own failings, his administration clung to an outmoded form of government. Ironically, Allègre, who never ceased to attack the unhealthy power exerted by French research mandarins and technocrats, was a pure product of this system. Power over a wide range of strategies and decisions was centralized at the ministry in the hands of small clans of powerful autocratic individuals and lobbies. Such overly top-down approaches to government are problematic, if not obsolete, in this fast-moving scientific era, where appropriate competences are widely distributed. Allègre adopted the role of a mainframe computer, and by undermining the autonomy of the research agencies, and by cutting himself off from them, he denied himself the power and speed of a network.

All the more frustrating, therefore, that the problems acknowledged as priorities by researchers are virtually identical to those that most troubled Allègre: the lack of independence of young scientists is one, the bloated system of evaluation another. And then there is the long-recognized rigidity of a system where some scientists enjoy full-time posts in the research agencies while their colleagues in the universities — often young scientists in their prime — struggle to find time for research alongside excessive teaching loads. But research agencies are reforming themselves, and while tough decisions will always be needed in some situations, an astute minister would realize that, in many areas, the agencies and the research community are badly needed allies. Schwartzberg will need to move swiftly to rebuild confidence in the research community, and seek to bring about negotiated change.

But he should also do a better job than his predecessor, and ministers in other countries for that matter, at the international level, where there are glaring opportunities in coordinating and planning sustained and cost-effective support for the second-generation Internet, international bioinformatics facilities, and other costly research facilities that serve large and diverse communities. As the recent funding fiasco at the European Bioinformatics Institute shows, insufficient national government attention has been paid to the crucial issue of European infrastructure. And whatever the merits of the recent French decision to contribute to building a new synchrotron in Britain, it was not the coherent pro-European exercise it was billed to be, but a hastily arranged cost-saving exercise.

France takes over the rolling presidency of the European Union in July. The biggest single opportunity facing Schwartzberg — and his counterparts in other European countries — is to make progress in the long-neglected challenge of building a scientifically coherent Europe in earnest.





WormBase
Internet site plans
to round up
nematode data
p425



Spanish jobs
New structure for
university
recruitment
p425



Shell collection
Giant success for
tortoise rescue
programme
p426



Bio-systems
Sharing simple
thoughts may solve
complex problems
p427

Safety concerns force NASA to condemn satellite to fiery end

Washington

The US space agency NASA last week chose safety over science and decided to send the nine-year-old Compton Gamma Ray Observatory to an early death.

The 17-tonne satellite suffered a gyroscope failure in December that compromised its ability to maintain a safe orbit (see *Nature* **403**, 232; 2000). It will be manoeuvred to burn up in the atmosphere over the Pacific Ocean on about 3 June.

Engineers at the Goddard Space Flight Center in Maryland, which operates the spacecraft, had found a way to control it even if all three onboard gyroscopes were to fail. But the new technique was estimated to be eight times less reliable than manoeuvring the satellite with gyros.

NASA fears that if the satellite fell uncontrolled, it would drop pieces weighing a tonne or more in an unpredictable path that could pass over populated areas. It could come down within three years, and the chance of a human fatality would be one in a thousand.

That was too much for agency science chief Edward Weiler, who made the decision to bring down the spacecraft. "I simply cannot justify more science in a mission that's already met its prime scientific objectives by increasing the risk of loss of human life by a factor of eight," he said at a press conference last week. Guiding the re-entry increases the safety factor dramatically, so that the risk of a fatality becomes one in four million.

Weiler insisted that cost was not a factor in his decision, as operation of the observatory, now running in a reduced mode, has been cut to less than \$4 million a year.

"We will lose science, there is no question about it," Weiler admitted. Compton project scientist Neil Gehrels said he was "personally profoundly disappointed" in the final verdict.

"The scientists were certainly pushing hard to see if we could eke out a few more months or years," said Gehrels. "But I respect the difficulty of this decision and the importance of safety."

No other spaceborne gamma ray telescope has Compton's all-sky view, and it is still discovering the enigmatic Gamma Ray Bursts (GRBs) at the rate of one a day. But



In a spin: gyroscope failure has doomed the Compton Observatory, released (above) in 1991.

astronomers will soon have a new spacecraft specially designed to study GRBs.

The low-cost HETE-2 satellite, which was to have been launched earlier this year (see *Nature* **403**, 583; 2000) is expected to reach orbit in June or July.

HETE is expected to find no more than

one to three GRBs a week, and its field of view will be limited. But, says principal investigator George Ricker of the Massachusetts Institute of Technology, ground-based observers should be better able to follow up rapidly on burst detections, as all HETE's observations will be in the night-time sky.

HETE's summer Pegasus launch replaces another small science mission, the High Energy Solar Spectroscopic Imager (HESSI), which last week fell victim to the latest mishap at NASA's Jet Propulsion Laboratory. A routine prelaunch vibration test went awry, and the spacecraft was briefly shaken with ten times the force it was supposed to receive, damaging its solar arrays and setting back its launch date by at least six months.

"It was a standard test on a standard facility — just bad luck," says Peter Harvey, HESSI project manager at the University of California at Berkeley. NASA will form a committee to determine the cause of the accident — an all-too-familiar exercise at the space agency these days. **Tony Reichhardt**

Software error 'caused Mars lander crash'

Washington

The Mars Polar Lander that disappeared without a trace last December probably crashed into the planet due to a software error, according to a report to be released this week. A committee chaired by John Casani, former chief engineer at the Jet Propulsion Laboratory, which managed the doomed spacecraft, identified the programming error after examining several possible accident causes.

The lander's legs had micro-switches that were to have signalled onboard engines to shut down when the legs made contact with the surface. But the switches sensed a jolt when the landing legs unfolded during the descent

through the Martian atmosphere, and faulty software failed to tell the difference between the jolt and a real landing. As a result, the engines shut off prematurely, and the lander fell the rest of the way to the surface with no retro-rockets to soften the blow.

A second panel of outside experts, chaired by former aerospace executive Thomas Young, identified more generic problems, as have other reports issued in recent weeks, which have criticized NASA for poor management and underfunding of missions ranging from the space shuttle to the Mars Climate Orbiter that also failed last autumn.

Tony Spear, a former manager of the successful Mars Pathfinder

project who reviewed NASA's 'better, faster, cheaper' missions for its administrator Daniel Goldin, said at a recent NASA Advisory Council meeting that "we should slow the rate of missions until we understand how to do them."

NASA and JPL are already planning to overhaul the Mars programme as a result of the reports, according to sources. The changes are likely to include more intense supervision of the programme at NASA headquarters, less emphasis on sample return as the main focus of the Mars effort, and using small missions to conduct a reconnaissance of the Martian surface before sending large landers.

T. R.

German researchers seek legal backing for stem cell work

Munich

German scientists were due to speak out at a public seminar this week against what they see as a legal double standard which allows them — at least in principle — to carry out research using human stem-cell lines produced commercially in other countries, but not to create their own.

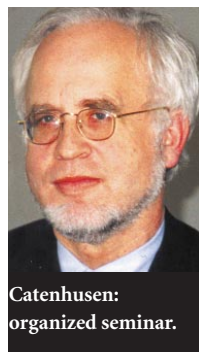
Germany's law on reproductive medicine, one of the most restrictive in Europe, bans the extraction of stem cells from a human embryo. Such cells have the potential to develop into any type of tissue, and could eventually be used to grow organs for transplantation.

But they also have the potential to grow into a complete human being, and the German scientific community has not previously pressed for a change in the law. "But science has moved on such a lot in the past year, and many think it's time for some details of the embryo protection law to be reconsidered," says Anna Wobus, a mouse stem-cell researcher at the Institute for Plant Genetics at Gatersleben.

Her view is shared by Wolf-Michael Catenhusen, social democrat secretary of state for the federal research ministry. He has organized a 'status seminar' in Berlin this week where scientists will outline the scientific possibilities and ethical limitations of human stem-cell research. The health ministry, run by members of

the Green party, is supporting the debate by holding a three-day seminar in May on all aspects of reproductive medicine. The ministry, which is responsible for proposing changes to the embryo protection law, is said to be less enthusiastic about the scientific possibilities, and few expect it to act during the current legislative period, which has three years to run.

This is not necessarily a bad thing, says Henning Beier, professor of embryology at the University of Aachen, and political adviser on reproductive medicine, because opening up a parliamentary debate could result in the law being tightened. Beier is one of many German scientists who feel uneasy that Germany can benefit from stem-cell research done in other countries while maintaining a 'moral superiority' by not doing such research itself. **Alison Abbott**



Catenhusen: organized seminar.

Balance is returning after US biotechnology shares scare

Washington

Biotechnology stocks recovered slightly last week from their sharp drop in value the week before. Many analysts suggested that the earlier call by US President Bill Clinton and British Prime Minister Tony Blair for the open sharing of raw gene-sequence data (see *Nature* 404, 324–325; 2000) was merely the trigger for a massive sell-off of overvalued shares that was already poised to happen.

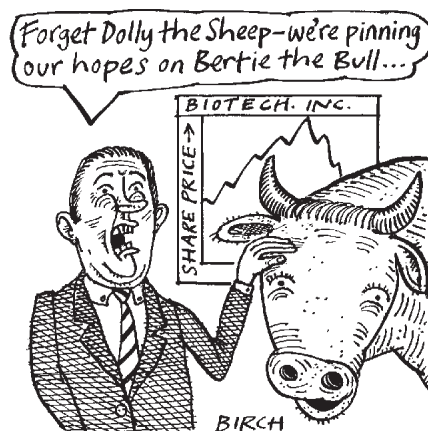
The analysts argued that the sector had become overvalued since December, when a stampede of apparently indiscriminating investors had poured money into biotechnology and genomics companies on the heels of a number of headline-grabbing announcements. These included the publication of the full sequence of chromosome 22 (see *Nature* 402, 447–448; 1999), and the news that Celera Genomics expects to finish sequencing the human genome this year.

Even if Clinton and Blair had not made their 14 March announcement, according to several analysts, some other event perceived as 'bad news' by investors would have prompted a similar sell-off soon. "The markets were looking for a reason to correct in biotech," says Eric Schmidt, a molecular biologist who now works as senior biotechnology analyst with S. G. Cowen Securities, a New York-based company focused on technology and health-care research. "By any measure the valuations were stretched."

"It was a speculative bubble," adds Karen Johnson Grunst, a health-care stock analyst at Banc One Investment Advisors, a large unit-trust manager in Columbus, Ohio. "People were investing in these stocks who didn't understand the fundamental stories — and more importantly, didn't want to."

Even though the Clinton–Blair joint statement emphasized the importance of patent protection for "gene-based inventions", many investors, apparently confused by erroneous media reports alluding to an incipient ban on gene patents, fled "in a blind panic", as one analyst put it.

By last Friday, the NASDAQ biotech index had fallen by 15 per cent since 14 March, registering a 30 per cent decrease since its post-December peak on 7 March. High-profile biotech shares have taken greater hits than the index as a whole, closing at strikingly lower values than those at which they opened on 14 March: Celera closed at \$122.5, down 31 per cent; Human Genome Sciences closed at \$97.81, down 37 per cent; Incyte closed at \$107.38, down 46 per cent; Millennium Pharmaceuticals closed at \$139.63, down 41 per cent; and Affymetrix



closed at \$155.75, down 35 per cent.

Further volatility and perhaps more falls in biotechnology shares are expected in the coming months. But analysts predict that solid companies should ultimately recover their lost ground. Indeed, many of the shares listed above increased in value between Monday and Friday last week; the NASDAQ index rose from its nadir of 1,087 on Monday to finish the week at 1,137.

William Haseltine, the chief executive officer of Human Genome Sciences in Rockville, Maryland, criticized the way the sell-off had taken place across the whole biotechnology sector. "Investors should look at the underlying companies, not the sector. I believe that if you look closely at our company, you will find a company worth investing in."

Indeed, some analysts urged clients to buy into what they now feel are undervalued companies that remain excellent long-term bets. Michael King, an analyst at Robertson Stephens, a San Francisco-based brokerage house, upgraded his recommendations from 'buy' to 'strong buy' on stocks including Affymetrix and Millennium Pharmaceuticals. The sell-off "does nothing to diminish our enthusiasm for the genomics sector nor the transforming technology contained therein," he says.

Some, however, are less bullish. Grunst, who co-manages \$1.15 billion in unit-trust health-care holdings, urged her company to replace pharmaceutical shares with biotechs last autumn. But in January and February she retreated from what she thought was becoming an overvalued sector, cutting her biotech holdings by a third and getting out of companies such as Biogen and Chiron altogether. Grunst says that she is now putting her money on companies that form the 'bricks and mortar' infrastructure behind biotechnology, along with selected product-oriented companies. **Meredith Wadman**

The worm's turn to claim Internet fame

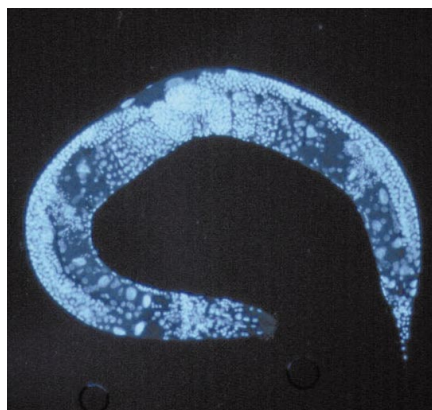
Paris

Caenorhabditis elegans, the favourite worm of developmental biology, now has its own Internet portal site. Wormbase, which can be found at <http://www.wormbase.org>, aims to bring together all available information on the millimetre-long soil nematode's genetics and biology.

Consisting of just 959 cells, the fates of which have been meticulously catalogued, the roundworm *C. elegans* is the leading model organism in developmental biology. Last year, it became the first complex organism to have its genome sequenced. Wormbase aims to integrate these data, and make them accessible to any biologist — not just those with expert knowledge of *C. elegans*.

"We want to make the database fun and easy to use, so that you don't have to be a roundworm insider to use it," says Paul Sternberg of the California Institute of Technology in Pasadena. Sternberg is leading the project, together with Lincoln Stein of the Cold Spring Harbor Laboratory on Long Island, New York, and Jean Thierry-Mieg of the Centre de Recherches de Biochimie Macromoléculaire in Montpellier.

John Spieth of the Washington University School of Medicine in St Louis and Richard Durbin from the Sanger Centre near Cambridge in England, representing the two



Browse a roundworm: visit www.wormbase.org

centres that sequenced the worm's genome, are also on board.

Until now, admits Sternberg, biologists outside the 1500-strong roundworm community may have found its obscure nomenclature off-putting. Few, for example, would have any idea that a cell called AB.alaaappr is the parent cell of AVDR, an important interneuron in the worm's ventral cord.

Wormbase users will require no such insider knowledge. "You will be able to come into Wormbase from a paper, a cell or a gene, or even a process," says Sternberg. For instance, the site will eventually allow users to

zoom in on photographs of the worm and see what cells such as AB.alaaappr look like, what they do, and which cells they are sitting next to. It will also document where and when each of the 19,000 genes in the worm's 97-million-base-pair genome is expressed.

"What is special about *C. elegans* is that we can see which genes are being expressed in which cell," says Sternberg, "and that we know what each of the cells does, and how they relate to each other." Ultimately, all the information should be available for a series of stages in the worm's development.

Wormbase is an example of a 'vertical' database that integrates many types of information about a single organism. Its model is Flybase, which serves researchers working on the fruitfly *Drosophila*. Sternberg believes that having the worm genome sequence early on was in some ways a disadvantage, as the focus on sequencing led to neglect of collating the literature on the worm's wider biology. The *Drosophila* community had to wait until earlier this month to get its genome, and wisely spent the waiting time in building up the Flybase portal, says Sternberg.

Another challenge for Wormbase will be interfacing its vertical information with the larger 'horizontal' databases, such as the GenBank sequence database, which span many species.

Declan Butler

Radical changes urged for Spanish universities

Barcelona

In an attempt to bring the structure and operating procedures of Spanish universities in line with those of other advanced nations, major reforms aimed at increasing the quality of scientific research and education at such institutions have been proposed in a report released in Madrid last week.

Entitled 'University 2000', the report has been drawn up by Josep Bricall, former president of the Council of European Rectors, at the request of the Council of Rectors of the Spanish Universities (CRUE).

Saturnino de la Plaza, CRUE's president, describes the proposals in the report as "unprecedented". In particular, the report recommends a complete revision of selection procedures for tenured university staff. It admits that this results from criticism of current recruitment procedures reflected in the large number of recent lawsuits which usually "base their claim on allegations of favouritism" (see *Nature* 400, 203, 401 & 419; 1999).

The report points out that, when the current higher-education law that confers increased independence on universities

came into force in 1983, it was initially thought that the high level of complaints over unfairness in the new recruitment system adopted at the time would be transitory. But even after 17 years, there are still a worryingly high number of legal challenges to selection procedures.

The new report now proposes that, once a post for an associate or full professor has been advertised, a panel of experts with no link to the university should make a judgement on the quality of the professional activities of the applicants, including their research, teaching and other activities.

The committee would list the candidates individually according to the results of this evaluation. A second panel, drawn from individuals at the institution advertising the post — but not from those dealing with the same professional tasks as those of the applicants — would select the successful candidate based on this external evaluation.

The report says that this system would improve the selection of professors because evaluation would not be based on a series of tests (or exams) assessed and voted on at the time of selection but instead on the

previous activities of the candidates.

CRUE, in its report, also recommends the creation of a nationwide independent Agency of Accreditation aimed at monitoring the quality of assessments being made at different universities (see *Nature* 404, 211; 2000). This agency would be staffed by individuals from the academic and professional worlds, both in Spain and abroad. The report says that it should make its results public, and that these should take the form of recommendations to the government.

The report's proposals have been widely welcomed in the academic community. Jordi Camí, for example, professor of pharmacology of the Universitat Pompeu Fabra in Barcelona, says that it is likely to boost competition between university researchers, as well as between universities. He is particularly enthusiastic about the novel system for selecting professors that the report proposes.

As far as research is concerned, the report makes a series of recommendations that are in agreement with the recently approved National Plan on Scientific



Building for the future: the new report suggests major changes for university tenure.

► **Research, Development and Technological Innovation** (see *Nature* 400, 393; 1999). It proposes in particular that there should be increased support for basic research and that links between universities and industry should be boosted.

In order to increase human resources in research and development, it says that a new permanent post of research professor should be created. The occupant of such a post would be employed directly by the universities and would address the research needs of a particular department as well as carry out PhD teaching activities.

A revised version of the report is likely to be presented to the Parliament in September, after comments received during a period of consultation have been taken into account.

The move no doubt reflects the high dissatisfaction level expressed by researchers and professors at Spanish universities. Such concerns appear to be sorted out, at least in part, in the CRUE report.

Raul Villar, rector of the Autonomous University of Madrid, says that the report should be considered as a consultation document which is now open to comment from academics and other social groups.

He points out that, although other countries may have adopted different recruitment procedures, those that are proposed in the report seem the most appropriate for Spain, as the report respects the independence of universities from the government.

"The report has to be strongly taken into account," says Juan Carlos Guerra-Zunzunegui, the parliamentary spokesman on education for the ruling Popular Party. He adds that he is keen to encourage a broad debate on university recruitment.

Xavier Bosch

Modified reforms end two years of deadlock at CNRS

Paris

Amid talk last week of the imminent dismissal of France's controversial research minister, Claude Allègre (see page 430), some of his suggested reforms for the Centre National de la Recherche Scientifique (CNRS) were given provisional approval — after two years of deadlock.

The reforms give the CNRS, Europe's largest public research agency, an internal administrative overhaul that has been backed by virtually everyone at CNRS. This includes the scientific unions that were bitterly opposed to Allègre when he tabled his initial reform proposals in 1998.

The most important changes approved by the agency's administrative council last week were welcomed by many scientists as bringing more flexibility and independence. These reforms include trimming the size of the scientific council — which defines the agency's scientific blueprint — and opening it to foreign members. The council will be elected internally, its members will be named, and it will choose its own president.

Another major change involves redefining the role of the president of CNRS, who will act as strategic director of the establishment and be responsible for relations with the government and universities. The director

general will oversee the day-to-day running of the institution. The CNRS will also have an external evaluation committee made up of French and foreign scientists.

While incorporating some of Allègre's original suggestions, the reforms have been stripped of the most controversial elements, such as reducing the influence of the National Committee for Scientific Research. This body acts as the 'parliament' of the country's scientists, and is involved in evaluating laboratories and administering recruitment. Allègre also stepped back from handing responsibility for much of the work carried out in CNRS laboratories to the universities.

Researchers have long criticized the CNRS for being too rigid. "It's a very important loosening up of the system," says Henri-Edouard Audier, a chemist at the École Polytechnique near Paris and a member of the board of a researchers' trade union, the Syndicat National des Chercheurs Scientifiques.

Audier believes the most important lesson to emerge from the two-year debate is that, once the CNRS had been given increased independence from the government, it had been able to agree on its own reforms without excessive government interference. "When there is consultation, the CNRS can modernize itself," says Audier. **Heather McCabe**

Giant tortoises come home

London

The one-thousandth giant tortoise (*Geochelone elephantopus hoodensis*) to be repatriated to its native Galapagos island of Española was released last week. This is a milestone in a breeding programme begun in 1963 to save a rare tortoise threatened by humans and introducted species.

Several centuries of exploitation had left a reduced and dispersed population that was not reproducing. The *G. e. hoodensis* project began with only 14 tortoises — all that could be found at the time. After being bred in captivity, the tortoises are reintroduced into their natural habitat (see right).

The project is led by the Charles Darwin Research Station on the Galapagos island of Santa Cruz, in conjunction with the Galapagos National Park. The research station now faces the challenge of dealing with many other threatened tortoise species. Officials say it will have succeeded when "humans are no longer needed to lead the tortoises of Española or other islands back home". **Natasha Loder**



CHARLES DARWIN RESEARCH STATION

Is starting simple the path to complexity?

Washington

The best way to understand complex biological systems may well be through integrating different scientific disciplines. But a panel of high-profile scientists agreed last week that the task of achieving this goal is itself inherently complex.

The panel discussion, held in Washington last week, deliberated the concept of interdisciplinary approaches to complex systems. The session came at the end of a three-day meeting on 'Biology: Challenges for the New Millennium', co-sponsored by the American Institute of Biological Sciences and the Smithsonian Institution.

Problems that would benefit from a complex-systems approach include understanding how elements in an ecosystem contribute to environmental change, and how proteins interact to produce particular end results.

"Do you start teaching synthesis in high school and at the undergraduate level?" asked Marvalee Wake, Chancellor's Professor of Integrative Biology at the University of California, Berkeley. "Is that the goal?"



Wilson: begin by understanding one component.

No, said Edward Wilson, honorary curator in entomology of Harvard University's Museum of Comparative Zoology and Pulitzer Prize-winning co-author of *The Ants* (Harvard, 1990). He advised starting with an understanding of a single component of a system — whether one insect in an ecosystem or one protein in a signal transduction pathway. "When you know enough about basic biology, then you are ready to attack complex systems," said Wilson.

"You have to be really good at something, and then know a little about a lot," echoed Daniel Janzen, professor of biology at the University of Pennsylvania.

Not everyone agreed. Richard Norgaard, president of the International Society for Ecological Economics, which encourages an integrative approach to ecology, argued that complex systems biology and ecology should be emphasized throughout education, not embraced "after your hair turns grey".

Gordon Orians, professor emeritus at the University of Washington, said that, if integration meant literacy in everything, science was setting itself up for "massive failure". Instead, he proposed that biologists should agree on the smallest set of shared knowledge necessary for 'bioliteracy', then be free to develop specialized expertise.

Gene Likens, professor of ecology at Yale and Rutgers, was pessimistic. Scientists would have to embrace complexity and do a better job of communicating with experts from other subdisciplines. "I don't even see us trying," he said.

Paul Smaglik

India and US bring collaboration out of the freezer

New Delhi

Ragunath Mashelkar, head of India's Council of Scientific and Industrial Research, describes collaboration with the United States as a "frozen chicken" that, as a result of political tensions between the countries, has been left in the freezer for nine years.

Last week, the two countries agreed to set up a joint forum to promote greater interaction between government research institutes, universities and private industry.

Despite its relative lack of funding and an uncertain agenda, Indian officials say that the creation of the forum marks an important step in renewing scientific relations between the two countries. "The forum has brought the chicken out of the freezer for warming up," says Mashelkar.

The agreement was signed by US Secretary of State Madeleine Albright and India's science minister Murli Manohar Joshi during President Bill Clinton's visit to New Delhi. Describing it as "a positive sign", Joshi said that he hoped the forum would give a "new direction" to effective scientific cooperation with the United States.

Neal Lane, Clinton's assistant on science and technology, said the forum would be beneficial to both countries. It is expected to start operating by the end of June.

Scientific cooperation between India and the United States peaked in the 1980s with the launching of some 250 collaborations.

It went downhill in the 1990s, when the United States wanted India to sign the Paris Convention and provide patent protection for products developed on joint programmes. India at first refused to amend its law on intellectual property rights (IPR) allowing patents on processes but not products.

Cooperation was virtually halted by Washington's wish to see India's nuclear missile programme ended, and by the nuclear tests that India carried out two years ago. "The forum is a sign that the door is opening again," says Valangiman Ramamurthi, secretary to the Department of Science and Technology (DST).

But Indian scientists, who had been hoping for more from Clinton — in particular a lifting of the sanctions applied to some 150 Indian research institutes and enterprises after the nuclear tests — were disappointed. "Other than the forum, I can't think of any major [scientific event] during the president's visit," Ramamurthi admits.

The forum was first proposed three years ago but was put in cold storage over the IPR issue. Last year India did sign the Paris Convention, and an amended patents bill, meeting most of the US demands, awaits a vote in parliament. The forum agreement was signed without an IPR clause that the United States had requested.

"Relaxing IPR means nothing," said an official accompanying Clinton. "Besides, the

political environment has improved." But Ramamurthi described the US decision to drop its demand for a clause on IPR as "a significant achievement for us".

The forum will be run by a governing board of seven members from each country. According to A. P. Kulshreshtha, head of the international division in the science ministry, it will commission studies and reports on cutting-edge technologies, and promote collaborative projects. The conference did not identify specific areas for collaboration; this will be done in a follow-up meeting to which industries from both countries will be invited.

Meanwhile, many Indian scientists continue to protest against sanctions on their laboratories. "The right environment should be created for cooperative research," physicist M. G. K. Menon, former science minister, told an Indo-US round-table conference at Hyderabad on 24 March. He described the sanctions as a "thorn" in the relationship between the two countries.

Questions remain over how the forum is likely to function without significant extra funding. The US is not putting any fresh money into the forum, only transferring unspent money — equivalent to \$7 million — from the now extinct US-India Fund. The forum would operate on the annual interest of \$700,000, with a matching grant from the Indian government.

K. S. Jayaraman

Chemists 'volunteered for nerve gas tests'

London

In a move that would bring instant disapproval from modern health and safety inspectors, former members of the chemistry department at Cambridge University have admitted that they tested nerve gases and other chemical warfare agents on themselves during secret World War II experiments.

On one occasion, chemist Fred Pattison went blind for ten days when his dosage was too high. "I thought I was permanently blinded," he now says.

In the early years of the war, when Britain feared that it would come under chemical attack, the Ministry of Supply arranged for several university chemistry departments to provide rapid response units. Staff of the chemical agents research team at Cambridge University were also encouraged to work on a number of potential warfare agents as a possible form of retaliation.

In the early 1940s, these researchers routinely tested new compounds on themselves. Staff were asked and "could have refused" to do the experiments, which were carefully monitored, says Pattison. But, he adds, everyone accepted the risks, as they saw their work as part of the 'war effort' and of national importance. "I remember being ushered



Chemical reactors: the Cambridge chemical agents research team in 1944. Fred Pattison (back row, second from left) was blinded for ten days.

into the gas chamber for 10-minute exposures, armed with paper and pencil to note down any symptoms," he says.

The human tests were intended to establish symptoms of exposure to the gases at sub-lethal doses. At that time, the researchers did not know the organophosphate compounds they were working on were nerve gases; only that they were extremely toxic.

"Everybody had a go," says John Ilett, another of the ex-Cambridge chemists. "I

can remember not being able to read anything for a couple of days." Ilett says as far as he knows nobody suffered long-term effects.

Alistair Hay, a senior lecturer in chemical pathology at the University of Leeds, is not surprised by the admissions of the Cambridge chemists. "There is a long history of scientists being involved in tests at Porton Down [in Wiltshire]," he says. After the war, Porton Down turned to volunteers from the armed forces.

Bernard Kilby worked in both the chemistry and pathology units at Cambridge and oversaw many of the human and animal tests. He says it was suggested that the units “do something useful” while they were waiting for a gas attack. The Cambridge unit was allocated lead and fluorine compounds, and tried to develop new toxic compounds based on those identified during the inter-war years. Kilby worked on making the fluorine analogue of mustard gas — widely used in the first world war.

When the Cambridge team discovered successful new substances, such as diisopropyl fluorophosphate (DFP), these were developed at the Ministry of Defence's Porton Down research centre. DFP went on to be manufactured on a large scale by the United States during the war. Gases were never used during the second world war, although once it was over, chemists realized that Germany had developed chemical agents much further than they had anticipated.

One of the chemists involved in the Cambridge experiments says he shudders to think what an ethics committee would say today about the research. But Ilett says they trusted the pathology lab's calculations on body weight. “It was war time. Everyone was taking risks. You were not offering yourself as sacrifice, but it took away the boredom of the war years in Cambridge”.

Natasha Loder

Cuts may force cancellation of Austrian research grants

Munich

Austria's new coalition government has slashed the annual budget of the Austrian Science Fund (FWF), the country's main funding agency for basic research, by a quarter, forcing the agency to consider cancelling two of the six rounds of grants awards that it makes every year.

The cuts are part of a broad cost-cutting budget introduced by the government. But what has particularly upset researchers is the fact that the cuts to the science fund are significantly higher than those being imposed on other areas of public spending.

Many have reacted with dismay to this decision, which means that no money for funding new projects will be available until after the summer. “This is absolutely unacceptable,” says Gunther Kreil, head of the department of biochemistry at the Austrian Academy of Sciences' Institute for Molecular Biology (IMB) in Salzburg.

The IMB employs about 20 PhD students and postdoctoral researchers on FWF grants, part of the 1,300 supported in the whole of Austria in this way. Although existing jobs

and projects may not be threatened in the immediate future, Kreil foresees bleak prospects for the next wave of researchers.

“It will become immensely difficult, even for very gifted young scientists, to get grant money,” he says, expressing particular concern at the impact that the move could have in competitive areas such as molecular biology. “It is a shame that this comes at a time when there is an abundance of excellent research ideas.”

Arnold Schmidt, the FWF's president, has called the cuts “a sheer disaster” for all basic research in Austria. He says that he is particularly disappointed that pre-election promises to increase public spending on research significantly over the next few years have apparently been abandoned.

With a mere 1.6 per cent of the gross national product being spent on research, Austria has never been the land of plenty for scientists, although some recent initiatives — such as the creation of a new Genome Research Centre in Vienna (see *Nature* 401, 630; 1999) — have brightened the general picture.

Quirin Schiermeier

Young UK scientists fear a bleak future in the academic world

London Three new reports on the situation of British research students paint a gloomy picture of a community worried about salary and career prospects in science. Two, published today by the Wellcome Trust, reinforce the widely held view that many young scientists are leaving the academic world because they have been put off by poor salaries and career structures.

The reports point out that those who have benefited from generous PhD stipends provided by the trust subsequently experience a drop in take-home pay when taking on most academic postdoctoral positions. Many of those who had earlier been supported by the trust expressed "considerable disillusionment" with the academic world. The reports are available at <http://www.wellcome.ac.uk/students>.

Meanwhile the pressure group Save British Science has produced its own report on issues concerning science students in UK higher education. A survey of research students showed that many felt the PhD stipend of around £6,500 (US\$10,300) a year to be too small. Students were also unanimous that at no stage was enough information provided to enable them to make good career choices.

European Union agrees to 'benchmark' timetable

Lisbon Leaders of the European Union agreed on a timetable for introducing a benchmarking system for member states' research policies at their summit in Lisbon last week. The system was proposed by research commissioner Philippe Busquin (see *Nature* 403, 696; 2000). Indicators for assessing performance in different fields, especially the development of human resources, will be ready by June.

Ministers also agreed to establish a 'European innovation scoreboard' by June 2001. In addition, up to 16.5 billion euros (US\$16 billion) will be made available in concessionary loans by the European Investment Bank to universities, schools, research organizations and companies to foster research and development in information technology. This will allow projects with a total value of 50 billion euros to proceed.

Scientists support GM food technology

Washington More than 1900 scientists, including two Nobel laureates, have signed a declaration in support of agricultural biotechnology that was released last week as protesters converged on a Biotechnology Industry

Organization conference in Boston. The declaration says that signatories support "the use of recombinant DNA as a potent tool for the achievement of a productive and sustainable agricultural system".

The organizers of the declaration, which can be found at www.agbioworld.org, began collecting signatures via the Internet in mid-January. In the first three days, they gathered the names of 600 scientists in academia and industry. "I'm very heartened," says C. S. Prakash, a professor of plant molecular genetics at Tuskegee University in Tuskegee, Alabama, who first conceived of the project to coincide with January's biosafety protocol negotiations in Montreal.

Human genome project is 'on track' for completion

Washington The publicly funded Human Genome Project has passed the two-billion-base-pair mark, and is on track towards completing a rough draft of the human genome in June, Francis Collins, director of the US National Human Genome Research Institute, announced in Boston this week.

Both the public project and the competing private version have established, passed and subsequently publicized meeting several markers. Celera Genomics, of Rockville, Maryland, announced last October that it had successfully sequenced a billion base pairs; the HGP said it reached the same goal a few weeks later, although the public project formally celebrated the achievement on 23 November (see *Nature* 402, 331; 1999).

Allègre ousted to calm angry teachers

Paris Earth scientist Claude Allègre has been dropped as minister of national education, research and technology, in a government reshuffle this week. Prime Minister Lionel Jospin, a lifelong friend of Allègre, made the decision after Allègre's unpopularity among schoolteachers reached crisis point, culminating in massive street demonstrations calling for his resignation.



Allègre: unpopularity reached crisis point.

Research and education are now split into separate ministries, each headed by politicians. The research minister will be Roger-Gérard Schwartzberg, a left-wing member of parliament and legal expert, who was secretary of state for education in the early 1980s. Jack Lang, a former minister of culture and one of France's most popular politicians, will be education minister.

Nobel prize-winner tipped as Taiwan's prime minister

Tokyo Taiwanese Nobel laureate Lee Yuan-tseh is believed to have been offered the post of prime minister last week by Taiwan's newly elected president Chen Shui-bian, leader of the Democratic Progressive Party. Although Lee would not comment on his intentions, he has been appointed to chair an ad hoc advisory committee that is drawing up a blueprint for the next government and criteria for selecting the new cabinet.

Lee, who recently resigned as president of Academia Sinica, won the Nobel Prize for chemistry in 1986 with Dudley Hershbach of the United States and Canadian John Polanyi for work on the dynamics of elementary chemical processes. A key supporter of Chen in the presidential elections, he is widely seen as being well placed to handle negotiations with the People's Republic of China.

Stirling to take over as synchrotron director

London William Stirling, an experimental physicist from the University of Liverpool, has been appointed as the next director general of the European Synchrotron Radiation Facility (ESRF) in Grenoble. His research interests include X-ray investigations of the magnetic structures and phase transitions of magnetic materials, areas that are among the core activities of the ESRF.

Stirling is well acquainted with the Grenoble research environment, having worked at the nearby Institut Laue Langevin between 1973 and 1987. He will succeed Yves Petroff from 1 January 2001, on a contract that lasts until 2006.

Ukrainian marine biologist takes up US post

London Ukrainian marine biologist Sergey Piontkovski arrived at Stony Brook University in New York state last week, after an ordeal that started last October, when he was accused of exporting state secrets from the Ukraine. Charges of illegal currency operations and 'organized crime' were withdrawn last month (see *Nature* 404, 10; 2000).

Friends and colleagues had organised a campaign to halt the prosecution and Piontkovski received high-level support from various non-governmental organisations.

Piontkovski says he is relieved that the "nightmare" of the investigation by the Ukrainian Security Service is over. But, with his wife and son, he went through what he describes as a "last farewell" from the security services two weeks ago. Shortly before their flight to New York departed, passport control officials detained the family for five hours. They were later released and allowed to leave the country.

Local farmers would be able to feed Africa if they were given the chance

Sir— Though overall food production in Africa has more than doubled in the past 30 years, it has declined by 14 per cent on a per capita basis. Many countries have now become dependent on foreign food aid and their increasing food bill constitutes a critical burden for future development. This has serious political implications, especially as a large untapped land-use potential exists in this part of the world¹.

Despite massive financial and technical assistance, the current trend is unlikely to be rapidly reversed unless the underlying reasons are better understood. Current explanations focus primarily on the adverse effects of population pressure, mismanagement of the land and poorly adapted farming practices. But in several west African countries, farmers' attitudes and incentives may be even more crucial, so appropriate government policies could really make a difference.

Agriculture in Africa is traditionally associated with low-resource production of food for local consumption. The vast population increase — on average 3–4 per cent per annum² — and the rapid development of urban centres has seriously disrupted this tradition, but has provided opportunities for a more market-oriented production. Yet, except in the immediate neighbourhood of the new centres, very few farmers have taken up this opportunity. In the more remote areas where most of the untapped potential is located, agricultural production has remained low and stagnates at subsistence level. This is not surprising because farmers in these poorly accessible areas have difficulty in transporting their goods and have no incentives to produce beyond their direct needs.

In Sierra Leone, for example, average yields for upland rice — the country's staple food — have dropped from 700 kg per hectare in 1978 to 516 kg per hectare in 1992. This is mostly explained by increasing population pressure: the length of the fallow period is shortened and the soil nutrient status depleted, further reducing yields. In rural areas, field observations indicate that instead of population pressure there is a shortage of labour, owing to the massive migration of younger people to the urban and mining centres. But this leads to similar results: land is reclaimed after a fallow period of, say, four to five years instead of the traditional 12 to 15, when forest regrowth would be denser and harder to clear.

At Freetown market in 1995, a bag of

imported (and subsidized) rice cost 280 leones (roughly \$26) compared with 320 leones for the locally produced rice. Though the latter is of high quality and suits local tastes, it cannot compete with the cheaper imported product.

The main reasons for this price difference are structural and beyond the farmers' control. For example, the poor interior road system increases transport costs, farmers are overdependent on middle men and there is no mechanism for setting a fair price. As a result, inland farmers are not interested in producing more than their families need and young people have no incentive to stay in rural areas.

The solution, nevertheless, is simple. If local farmers can compete properly with imported goods, they can take a share of the growing demand. This requires the government to support indigenous farmers by abandoning its subsidy policy for imported food products; to improve the interior road system, permitting easier transport of local products to the markets; and to support fair-price setting.

These proposals are similar to agricultural policies in western free-market economies. To the individual farmer, they will provide incentives for a higher return and should encourage better use of the land, ultimately improving farming practices. Gradually, producers will think of themselves as part of the internal economic market system rather than as poor subsistence farmers without any social status. At the national level, such policies will reduce the income gap between urban and rural areas and may reduce migration to the cities, as well as leading to a lower external food bill.

Willy H. Verheye

Fund for Scientific Research Flanders, Department of Geography, University of Gent, Krijgslaan 281, B-9000 Gent, Belgium

1. FAO World Soil Resources, an Exploratory Note on the FAO Soil Resources Map at 1: 25 Million Scale: FAO World Soil Resources Report 66 (FAO, Rome, 1991).
2. United Nations Development Programme Human Development Report (Oxford Univ. Press, Oxford–New York, 1999).

Ancient Chinese had their fingers on the pulse

Sir— In his Millennium Essay "Where do babies come from?" Roger Short calls William Harvey's seventeenth-century *De Motu Cordis et Sanguinis in Animalibus* "the first accurate description of the circulation of the blood".

Readers may be interested to know that such a description existed in the Chinese medical treatise *Su Wên*, written in the second century BC. It covered circulation of both blood and, in different channels, *qi* or *chhi* (roughly translated as 'vital energy',

with no equivalent in Western science). As science historian Joseph Needham² writes: "We can find statements such as this in the *Su Wên*, where Chhi Po says that 'the flow in the tract and channel runs on and on, and never stops; a ceaseless movement in an annular circuit...' Clearly the circulation of the blood and *chhi* was standard doctrine [in the second century BC]".

The heart's pumping function was also known. Needham writes: "In the *Lei Ching*, Chang Chieh-Pin wrote: 'The heart and pulse (*mo*) is not itself either *chhi* or blood (*hsüeh*), but rather it is the bellows of the *chhi* and the blood'. Chang published his compendium of medical physiology four years before *De Motu Cordis* appeared.

The ancient Chinese also wrote that 50 complete circulations of the bloodstream would take one day². Needham even suggests a mechanism by which they could have estimated the time required for one complete circulation, which, given the experimental spirit of the ancient Chinese, seems a reasonable conjecture. No concrete evidence of such experimentation has been found so far, probably because blood circulation was not of paramount interest to Chinese physicians compared with the circulation of the *chhi*.

Some seventeenth century European texts referred to Chinese descriptions of blood circulation. But since then, a misunderstanding has arisen that Harvey was the first to describe this process.

P.-L. Chau

Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, UK

1. *Nature* 403, 705 (2000).
2. Needham, J. & Lu, G. D. *Celestial Lancets*, 29–32 (Cambridge Univ. Press, Cambridge, 1980).

No pie in the sky, thanks

Sir— Geoffrey Cantor has overlooked the reason why many scientists perceive a conflict between science and religion, in his Millennium Essay "Fighting the wrong battle" (*Nature* 403, 831; 2000). Science is an enterprise whose proponents attempt to arrive at the truth by reasoning from experiment and observation. Religions rely on beliefs for which there is no support other than the dogmatic statements of sacred writings or clerics, and which are often imposed on the populace by force.

It is ludicrous to maintain that there need be no conflict between such fundamentally opposed activities. If the human race is to survive the next millennium, we need to face reality, not delude ourselves with fairy stories.

Brian Charlesworth

Institute for Cell, Animal and Population Biology, King's Buildings, West Mains Road, University of Edinburgh, Edinburgh EH9 3JT, UK

Scientist, social reformer, soldier, spy

From a Massachusetts farm to the court of England's George III, and beyond.

Count Rumford: The Extraordinary Life of a Scientific Genius

by G. I. Brown

Sutton: 1999. 182 pp. £12.99, \$21.95

John Meurig Thomas

Apart from being a formidable scientist and ingenious inventor, the American-born Count Rumford was also a soldier, spy, statesman, a brilliantly successful, if callous, social reformer, and the individual predominantly responsible for founding the Royal Institution of Great Britain. Born in 1753 in Massachusetts, of simple farmer stock, Rumford also founded the professorship named in his honour at Harvard University, and the Rumford prizes and medals of both the Royal Society, London, and the American Academy of Arts and Sciences, Cambridge, Massachusetts.

In quick succession, while in his middle teens, Rumford (who was christened Benjamin Thompson) tried his hand at commerce, medicine and schoolteaching; he also attended some courses at Harvard. In July 1772, he took up a position as schoolmaster in the town of Concord (formerly called Rumford, a corruption of the English town Romford), New Hampshire. Four months after arriving there, he married the wealthiest widow in the province. She was nearly 14 years his senior. He thereby became not only a gentleman farmer — with scope to indulge in natural philosophy — but also gained entry into the political and social circle surrounding the Royalist governor, John Wentworth.

Rumford became extremely unpopular in Concord because of his overt espousal of the American Tory cause in that turbulent period when the American colonies were struggling to free themselves from the British Crown. He abandoned his wife and young daughter in Concord and fled, first to Boston and then to London, where the self-assured 23-year-old major of the New Hampshire Militia soon became private secretary to the Secretary of the Colonists, Lord Germain. It was in Germain's country estate that he carried out impressive experiments on gunpowder, and where he measured the momentum exchange between a bullet and a ballistic pendulum. The Reverend G. E. Ellis, writing of this period of Thompson's life (in a *Memoir* published in Boston in 1871), says that he was already a courtier. "He manifested in early manhood the tastes, aptitudes, and cravings which prompt their possessor, however humbly born, and under whatever repression from surrounding influences, to



Right, Rumford, aged 30, and, above, James Gillray's cartoon, "The comfort of a Rumford stove".

push his way in the world by seeking the acquaintances and winning the patronage of his social superiors, who have favours and distinctions to bestow". How true. He courted favour with, and gained the admiration of, Sir Joseph Banks, the influential president of the Royal Society — Thompson was elected a fellow in 1779 — and he "extracted" a knighthood from King George III before settling into an 11-year stint as the supremo of the Elector of Bavaria, Karl Theodor, who in 1791 made him Count of the Holy Roman Empire.

All this and more are described in this compact and well-illustrated account of Rumford's numerous adventures, good and bad, by G. I. Brown, former chemistry master at Eton. Brown also deals concisely with the sequence of events that led to the creation of the Royal Institution. Rumford wrote extensively about his own achievements and aspirations. There are also early biographies by Ellis and Bence Jones in 1871, as well as two recent ones by the late Sanborn Brown of the Massachusetts Institute of Technology who thoroughly analysed Rumford's accomplishments. Brown has benefited greatly from Sanborn Brown's five-volume publication of *The Collected Works of Count Rumford* (Harvard University Press, 1970–71), and he has written an admirable summary of the man known to most scientists largely through his famous observations on the drilling of cannon in the arsenal at Munich, observations that were to

demolish the caloric theory of heat. But there is much more, scientifically and technologically, to Rumford than that important work. He was among the first to enunciate the now-popular mantra that fundamental research in a problem is a prerequisite to technological development.

Brown's book highlights how, during Rumford's mission-oriented efforts to improve the conditions of the wretched inmates of his House of Industry (essentially a workhouse) and of the Bavarian army, many scientific advances were made. These included the discovery of convection currents in liquids, an estimate of the mechanical equivalent of heat, an understanding of radiation loss from

surfaces and the creation of a whole host of gadgets, such as pressure-cookers, drip coffee-makers, efficient oil-lamps, stoves, kitchen ranges and the construction of a photometer to measure candle power.

We are also informed how badly Rumford treated his first wife (his daughter less so), how he came to marry, and separate acrimoniously from, the

widow of Antoine Lavoisier, and how outrageously he treated Thomas Garnett, the first professor of chemistry at the Royal Institution. He also records Rumford's many amorous exploits: there was a succession of mistresses and two illegitimate children.

My only serious criticism of the book is that, in summarizing the dazzling number of discoveries made by Michael Faraday at the Royal Institution, two of Faraday's principal achievements are not mentioned: the introduction of the concept of field, which set the stage for James Clerk Maxwell's endeavours in the 1850s, and the realization that matter and electricity are inextricably linked. That electricity comes in units is of far greater significance than the laws of electrolysis *per se*. Brown also quotes a tendentious example to support his contention that Rumford wrote badly. Yet Rumford could make his case beautifully in the written word. How about this sentence? "The ardour of my mind is so ungovernable that every object that interests me engages my whole attention and is pursued with a degree of indefatigable zeal which approaches to madness." Nevertheless, this is an enjoyable book about one of the great individualists of science. ■

John Meurig Thomas is at the Master's Lodge, Peterhouse, Cambridge CB2 1QY, UK.

FOGG ART MUSEUM, HARVARD UNIV.

Fingering the brain's past

Minds Behind the Brain: A History of the Pioneers and Their Discoveries

by Stanley Finger

Oxford University Press: 2000. 384 pp.

\$35, £24.99

Ian Glynn

Six years ago, Stanley Finger published his monumental *Origins of Neuroscience* (Oxford University Press), nearly two kilograms of double-column format, profusely illustrated and fully referenced, with a cast of some 900 involved in events extending over more than four millennia. Inevitably, with such a wide scope, much had to be omitted or condensed. Even the principal characters held the stage for too short a time to become fully rounded. The book had an encyclopaedic quality, and has, I suspect, been much more often dipped into repeatedly than read from cover to cover — hence the need for another and different book on the same general topic.

The immediate stimulus for *Minds Behind the Brain*, Finger tells us, came from the students who attended his lectures at Washington University, St Louis, Missouri. They wanted to know more about the brain pioneers “as real people”; they wanted to see more fully what led to their discoveries, and to understand “the ramifications of their insights ... knowing the year of a landmark and a ‘beard’ was not enough”. And Finger himself wanted space to look at the scientific literature in a social context. To achieve these aims he has had to prune his cast list drastically. He started, he tells us, by choosing a dozen highly influential players; but this figure waxed and waned, settling at 19, with a sizeable supporting cast who provide the background and illustrate the consequences of the achievements of the principal characters.

And the recipe works. Finger's erudition is remarkable. Even the characters we thought we knew about appear in a new light. René Descartes' conviction that animals lack consciousness seems even more remarkable when we learn not only that he had a pet dog, but also that he was extremely fond of it. We know about Galen's successes, so it is refreshing to hear of one of

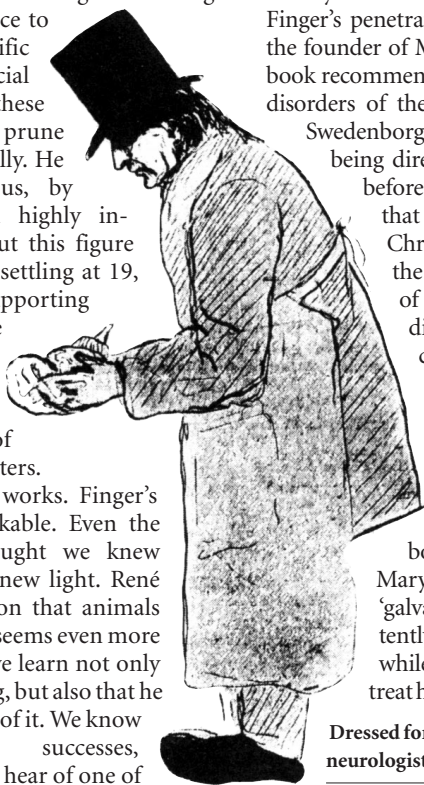
his failures — an attempt to increase the efficiency of leeches, by clipping their tails so that they could draw more blood, had to be abandoned when a sharp rise in leech mortality led to an increase in price. Those of us who have been raised on Michael Foster's *History of Physiology* have to revise (favourably) our views of both the character and achievements of the seventeenth-century anatomist and physician Thomas Willis. We know about Leyden jars, and we know that the Abbé Nollet, in eighteenth-century Paris, did party tricks with static electricity; but we may be surprised to learn that it was Nollet who coined the phrase ‘Leyden jar’, and that he once used one to make a 900-foot line of Carthusian monks jump en masse. Conversely, we might be astonished to discover that Franz-Joseph Gall, the founder of phrenology, never used the word phrenology.

But my favourite improbable story is of the physiologist Charles Sherrington as an immunologist. He and a colleague had been inoculating horses with small doses of diphtheria toxin in the hope that the animals would make a useful antitoxin. Hearing one night that his eight-year-old nephew had contracted diphtheria, he dashed down to Sussex with a flask of antiserum from one of the horses and, on being told by the family physician that the boy was expected to die within the next few hours, administered the antiserum. To the surprise of the physician, and to Sherrington's delight, the boy recovered within a day.

Even historical characters whom most of us would never have associated with the history of neuroscience are caught in Finger's penetrating spotlight. John Wesley, the founder of Methodism, wrote a popular book recommending electrical treatment for disorders of the nervous system. Emanuel

Swedenborg, after he had retired from being director of Swedish mines and before he experienced the visions that led to his reinterpretation of Christianity, was fascinated by the idea that different functions of the body were controlled by different areas of the cerebral cortex. Basing his arguments mainly on the observations of others, he reached original and important conclusions which, unfortunately, did not become widely known. In the next century, both Percy Bysshe Shelley and Mary Shelley were fascinated by ‘galvanism’; and Shelley inadvertently electrocuted the family cat while trying to use electricity to treat his sister's skin disorder.

Dressed for the job: nineteenth-century neurologist Jean-Martin Charcot at work.



Inevitably, the draconian slimming down of the cast means that the choice of those who are left must sometimes appear arbitrary. In discussing localization of function in the cerebral cortex, Finger mentions Swedenborg, Gall, Paul Broca, Carl Wernicke, James Ferrier and Eduard Hitzig, but says nothing about Salomon Henschen, Louis Verrey, Jules Déjerine or Harvey Cushing; Wilder Penfield is mentioned only very briefly. The discovery of chemical transmission at synapses is described at length, but the long history of the discovery of the nature of transmission along nerve fibres is lightly skimmed over. These are not serious criticisms. The book is designed as a series of fascinating excursions into the past, not an all-inclusive grand tour. Each of us may miss favourite characters and topics; but we are rewarded by the extra time that Finger can devote to the people and topics he chooses to write about.

Ian Glynn is at Trinity College, University of Cambridge, Cambridge CB2 1TQ, UK.

Before the great binary divide

Ingenious Pursuits: Building the Scientific Revolution

by Lisa Jardine

Doubleday: 1999. 480 pp. \$35, £25

Bruce Mazlish

The notion of ‘two cultures’ — art and science — has taken on mythical status. The great binary divide still possesses our imagination in spite of an original misunderstanding — C. P. Snow was less concerned with the break between science and literature than with a policy implication, that the lack of training in science could hamper economic development. The divide continues even though this leaves no room for a third culture, that of the social sciences, and despite the fact that the rise of multiculturalism has claimed to make the two-culture formula obsolete.

There are numerous reasons for the persistence of the common idea of two cultures. Many natural scientists are happy to see their subjects reaffirmed as being separate, that is, not contaminated by subjectivity from the soft humanities. Many humanists welcome the divorce as underlining their difference from the cold, calculating, ‘inhuman’ scientists. Either faction could appeal to much of the history of science as it was written in the period prior to the past few decades. Indeed, until recently, and still often today in practice, the natural and the human sciences have been treated, without further discussion, in a more or less compartmentalized manner. Snow, in contrast, at least put the subject on the table,

albeit in rather dramatic fashion. It is to the credit of Lisa Jardine, professor of Renaissance studies at the University of London, that her latest book demonstrates (although does not formally argue) the continuity, if not unity, of the two cultures.

Jardine's subject is seventeenth-century, everyday scientific life. Here there was no opposition between natural science and art and literature; rather, they were entwined. Robert Hooke had studied portrait painting before he published *Micrographia*, with its wonderful illustrations. Christopher Wren practised architecture with an eye to his buildings, such as St Paul's cathedral, being used for astronomical observations. William Harvey's anatomy was intimately intertwined with its representations in engraved plates. A look at the roster of the Royal Society, where a John Locke sat next to a Robert Boyle, confirms the lack of boundaries as we know them.

Jardine's canvas is broad. She ties together the 1680 comet with the construction of the Royal Observatory at Greenwich, and links both with the needs of war and the nation. Mapping is connected to the coming of clockwork, and the exact measurement of small intervals so essential to the development of science at large, and especially to the solution of the problem of longitude. Drainage, land reclamation, fortifications, the air pump, the transfusion of blood, the use of diving machines, the collection of specimens — all are to be found here inter-relating in sometimes obvious and other times strange ways. Much of what has been called the 'Scientific Revolution' is to be found here.

In many ways it is a coffee-table book (oddly enough, this is a phrase Jardine uses for Hooke's great work), for it is lavishly illustrated, with many coloured plates. It is primarily based on secondary work, as it should be, for it is a work of synthesis, placing mostly known material in new juxtapositions. Yet it rises towards originality in offering an "anthropology of science" that depicts the seventeenth-century sciences in ordinary, practical life, with practitioners clamouring for priority and notice, often filled with spite, eager for advancement. It is an application of Bruno Latour to an earlier period.

In her own life, Jardine exemplifies the personal aspect of scholarly or scientific work. As the daughter of Jacob Bronowski, whom many identify as the real source of the notion of the two cultures and its solution — certainly its exemplification — and to whom the book is dedicated, she carries on in his spirit in exemplary fashion.



An eye for detail: seventeenth-century Dutch artists used magnifying glasses for their still-life paintings, such as the one shown here of flowers with insects.

Her own starting point is the situation today, where the natural sciences are under attack for their nefarious consequences. Fittingly, the first sentence of her book refers to the publication in *Nature* of the 1997 report on the cloning of Dolly the sheep, and then goes on to the subsequent attacks on human hubris in seizing upon such God-like powers. The epilogue concerns *The Double Helix* (Penguin, 1999), and its account of scientific genius and personal ambition. Indirectly, then, Jardine's book is as much about contemporary science as it is about its seventeenth-century antecedents.

There are caveats. The illustrations are illuminating, but are rarely tied in significantly to the text. And why are the colour plates disconcertingly repeated in black and white (certainly running up the cost of the book)? Publishing data on cited material are either missing or hard to find. The chronology is occasionally confusing. The emphasis is on English material, with some attention to the Dutch and French, although the book's stated thesis is on the international nature of science.

The dark shadows of science are relegated to the far corners, only occasionally allowed to peep through with regard to vivisection or military affairs. This is a generally sunny picture of science (one, I confess, I myself lean towards, although I believe more attention should be paid to the dark side as well).

Putting such caveats aside, one must applaud Jardine's achievement. In the same

mode as her highly successful *Worldly Goods* (Papermac, 1997), she has sought to make scholarly research available to a broad public. In 1672, as Jardine tells us, Henry Oldenburg complimented Isaac Newton on his "ingenuity". I am moved to describe Jardine's own work as 'ingenious'.

Bruce Mazlish is in the History Faculty, Massachusetts Institute of Technology, Cambridge, Massachusetts 02138, USA.

Ill-served inventions

The Greatest Inventions of the Past 2000 Years

edited by John Brockman

Simon & Schuster: 2000. 192 pp. \$22

David Jones

An editor should be respected, even feared, maybe even disliked, by his contributors. His job is to dock their prolixities, unify their styles, and force them to conform to a consistent and clearly defined vision of the work in hand. It is no job for a democrat.

The trouble that can result from democratic, permissive editing is sadly illustrated by *The Greatest Inventions of the Past 2000 Years*. John Brockman invited a coterie of "today's leading thinkers" to pen — or rather to e-mail — their thoughts on this topic. They seem to have had a lot of fun with it, e-mailing chatty ideas back and forth. Brockman then assembled 109 of the resulting messages into this book.



Not surprisingly, it is a very mixed bag. Most scientists will have some instant nominees for great inventions: the lens, printing, the digital computer, artificial light, the gun, the contraceptive pill — and indeed, all of these are present. Other less obvious candidates, persuasively argued in these pages, include hay, the stirrup, paper, the mirror, distillation. There are some staggering omissions — mechanical power, radio. And the book is sadly flawed by many casual, bitty entries unworthy of the topic.

It seems clear that the thinkers were held to no clear remit. The entries include inventions physical (the battery), procedural (scientific method), social (the university), intellectual (arabic numerals) and even imaginary (the unknown benisons confidently expected to flow from genetic sequencing).

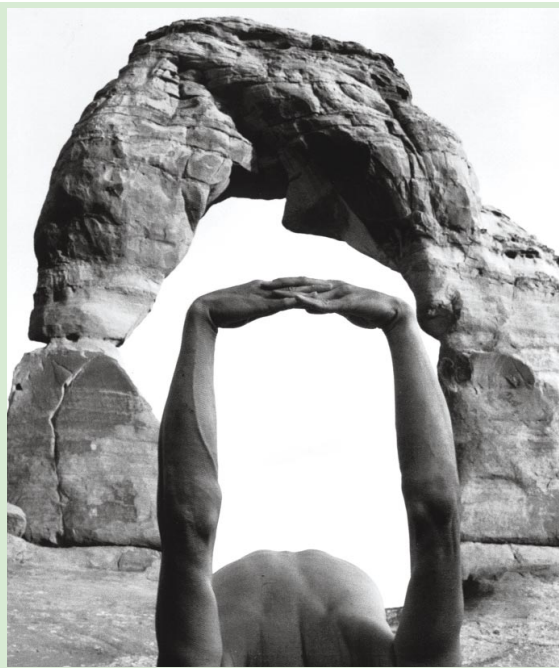
Neither has the editor demanded that his leading thinkers justify their title. Many, indeed, think intelligently about their field, consider the history of the past two millennia and what greatness means in an invention, before presenting an argument for the invention of their choice. But others are allowed to get away with a few scrappy, thoughtless sentences. One thinker champions “the construction of autonomous tools”, but in 34 lines gives no clue as to what this means, if anything, or whether it has yet been achieved. Another, in just four lines, champions the printing press to avoid upsetting the editor, and the Thermos bottle for no reason. Another nominates television, but remarks of this pervasive, socially transforming invention merely that it provides information, influences patterns of behaviour such as sex and crime, and has reduced the popularity of live performances. Another leading thinker, in the course of a long and rambling rant against the whole idea of invention, states (without explaining why it is relevant to the topic in hand) that he “doesn’t think he could hack ... assholes and complacent dopes living hundreds of years”.

The inclusion of such sloppy and demagogic language is an extreme instance of editorial neglect of duty. But the book as a whole suffers from the lack of a consistent voice. Many entries, for example, address conversational remarks to the editor, or to other contributors. To present such casual random chatter in book form is discourteous to readers outside the charmed circle. And, given the mighty computer power available to all concerned, the lack of any index or contents classification suggests further neglect. One invention celebrated in the book is the eraser in its many forms. It even features on the dust-jacket. It should have been wielded far more vigorously inside. ■

David Jones is in the Department of Chemistry, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK.

At one with nature

Arno Rafael Minkinen, a teacher of photography, has attempted to express the relationship of our body to the land it inhabits, through a collection of self-portraits. His naked body is depicted in a range of isolated settings, emphasizing its bond with the natural world. The collection, called *Body Land* (Smithsonian Institution Press, \$24.95, £15.95), includes *Double Delicate Arch*, shown here, taken in Moab, Utah.



A view of figures through the ages

The Human Body: Image and Emotion

by I. Philippe Comar

Thames & Hudson: 1999. 159 pp. £6.95 (pbk)

Ludmilla Jordanova

The Human Body is not a book about science, nor is it written by a scientist, yet it touches on scientific and medical questions at many points. The author is an artist with an interest in the representation of the human body. This is a huge field and, accordingly, the book covers a long time span, essentially the beginnings of art to the present day.

Philippe Comar includes diverse images from the expected anatomical drawings and nudes to carvings, photographs, Greek statues, wax models, drawings and various items of art that use the human figure in some way. The illustrations are of a high quality and, although some will be familiar to most readers with an interest in the body, many are not, and hence their reproduction in this accessible, low-cost form is to be welcomed.

The organization of the book is unusual and rather unsatisfying. Pages are crammed with pictures, the main text and quotations. Each image also has a caption which responds to it in some way rather than being purely factual. This is indicative of the book’s overall approach. It consists of five essays, which are meditations on themes such as “the outraged body” and “embodying desire”. Each one ranges freely in time and space, and their titles tell the reader nothing about their content.

At the end of the book, there is a short section comprising extracts from texts and

supplementary pictures. It is clear that it is partly to be read for pleasure and partly to be used as a compilation of sources and ideas, although it makes no claims to be systematic.

The question of how, over many centuries, the human body has been a central concern for fields we now differentiate as art, medicine and science is formidably complex. These intersections have become a fashionable subject, but here they are often dealt with superficially. This is not an intellectually deep book.

It originally formed part of a series developed by the French publishers Gallimard, and it is fully in keeping with the significance accorded to *haute vulgarisation* in France. When done well, books that are heavily illustrated, well written and informative perform an invaluable service, as long as readers know what they are getting.

Although not a scholarly work, it is thought-provoking and I came across only one topic, physiognomy, that I felt was presented misleadingly. Because of its association with racial classifications, this is always a difficult subject to handle well. A more intellectually rigorous account would have to pay considerable attention to the contexts in which specific views of the body were generated, which is all but impossible in works of this type. Hence Comar’s visual and verbal essays have their merits, as long as they are taken as being personal responses rather than authoritative scholarship. ■

The challenge of writing about the long-standing, intricate and fascinating intersections between the visual arts and knowledge of the human body, in a way that both does them justice and is accessible, remains.

Ludmilla Jordanova is at the School of World Art Studies and Museology, University of East Anglia, Norwich NR4 7TJ, UK.

The Big Bang and the genetic code

Gamow, a prankster and physicist, thought of them first

Gino Segrè

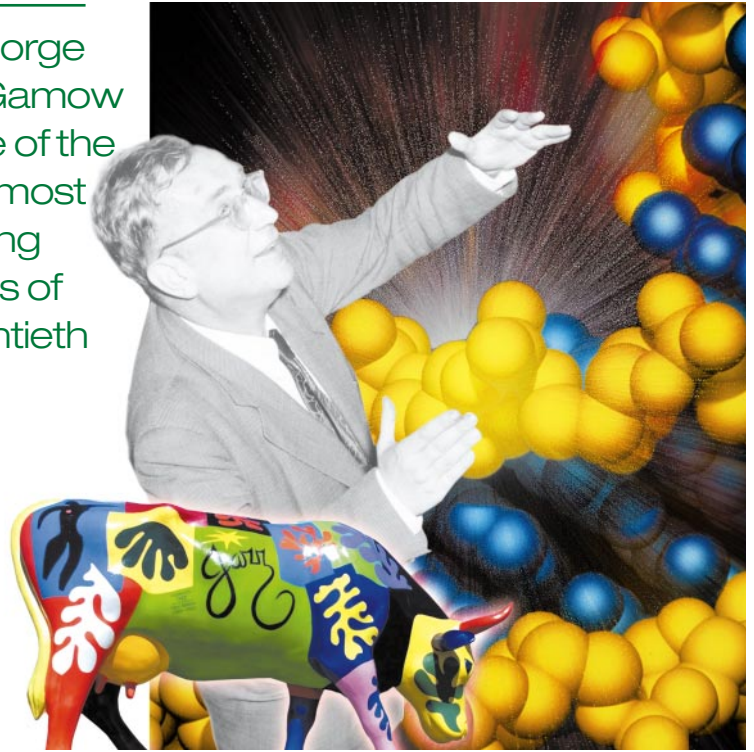
The Big Bang and the genetic code are probably the two scientific ideas that have most radically altered our view of the world in the twentieth century. The Big Bang seeks to explain how the Universe was created and how the primordial constituents came to be formed. The genetic code lays down the pattern for the formation of living material and the transmission of inherited characteristics.

Interestingly enough, the first serious theories of both these key ideas were formulated by the same man, a colourful, irresistibly playful physicist named George Gamow. He was born in Odessa in 1904, began his education there and then moved to St Petersburg. In 1928, Gamow demonstrated that nuclear alpha decay could be understood by quantum mechanical tunnelling. He also showed that inverse tunnelling could occur, that is, sufficiently energetic protons would penetrate the nuclear barrier; this stimulated Cockroft and Walton to build the first machine capable of artificially disintegrating nuclei. So began, in 1932, the world of experimental nuclear physics.

A 1940 paper by Gamow and Mario Schoenberg was the first in a subject we now call particle astrophysics. The two authors presciently speculated that neutrinos could play a role in the cooling of massive collapsing stars. They named the neutrino reaction the Urca process, after a well known Rio de Janeiro casino. This name might seem a strange choice, but not to Gamow, a legendary prankster who once submitted a paper to *Nature* in which he suggested that the Coriolis force might account for his observation that cows chewed clockwise in the Northern Hemisphere and counter-clockwise in the Southern Hemisphere.

In the 1940s Gamow began to attack, with his colleague Ralph Alpher, the problem of the origin of the chemical elements. Their first paper on the subject appeared in a 1948 issue of the *Physical Review*. At the last minute Gamow, liking the sound of 'alpha, beta, gamma', added his old friend Hans Bethe as middle author *in absentia* (Bethe went along with the joke, but the editors did not). Gamow and Alpher, with Robert Herman, then pursued the idea of an extremely hot neutron-dominated environment. They envisioned the neutrons decaying into protons, electrons and anti-neutrinos and, when the universe had cooled sufficiently, the neutrons and protons assembling heavier nuclei. They even estimated the photon

George Gamow was one of the wittiest, most interesting scientists of the twentieth century



Practical jokes or profound ideas: Gamow considered a breathtaking range of scientific questions.

background that would be necessary to account for nuclear abundances, suggesting a residual five-degree background radiation.

We now realize that their scheme was incorrect. The Universe began with roughly equal numbers of protons and neutrons. Collisions with electrons, positrons, neutrinos and anti-neutrinos are more important than neutron decay, and the absence of stable nuclei with atomic numbers of five and eight creates a barrier to further fabrication in the early Universe. Nevertheless Alpher, Gamow and Herman's work was the first serious attempt to discuss the observable consequences of a big bang and the basic framework was correct. Ironically, the term 'Big Bang' was coined by Fred Hoyle, an advocate of a steady-state model of the universe, to make fun of Gamow's efforts.

In 1953, Gamow read the famous Watson and Crick *Nature* Letter on the structure of DNA and immediately jumped to the conclusion that the DNA molecule could directly serve as a template for protein synthesis. Recognizing that there are twenty independent combinations of triplets that can be made out of four letters, he suggested that there are twenty relevant amino acids and the existence of a one-to-one correspondence between triplets and amino acids, the

backbones of proteins. In his autobiography, Crick remembers receiving a letter from Gamow suggesting this possibility, and then realizing that he and Watson had not even counted the number of amino acids. Crick, Gamow and Watson became good friends and Gamow even founded an RNA Tie Club, limited to twenty members, each one of whom had a special tie, provided by Gamow, with a specific amino acid in its design. The connection is not as direct as Gamow had envisioned: DNA makes RNA which makes amino acids and the code is not one-to-one, but Gamow's 1954 *Nature* Letter, "Possible relation between DNA and protein structures", was the first to appear on a subject which has changed our lives.

The DNA manuscript listed a Mr Tompkins as a co-author, but a vigilant editor removed his name. Gamow had written a series of popular books, in which a mythical Mr Tompkins explores the wonders of the world. Even without Mr Tompkins, Gamow was one of the wittiest, most interesting scientists of the twentieth century, a man with an extraordinarily quick and sure grasp of the great problems.

Gino Segrè is a professor of Astronomy and Physics at the University of Pennsylvania, 209 S. 33rd Street, Philadelphia, Pennsylvania 19104-6396, USA.

Eternal verities, eternal questions

How our ancestors reconciled themselves to eternal life

Ben Bova

It is difficult for us to understand the controversy that was caused early in the twenty-first century by the advent of human immortality. When, in the late 1990s, the Geron Corporation of California and the University of Texas Southwestern Medical Center announced that they had 'immortalized' human cell cultures, the general public hardly took any notice.

But when the Fossel Clinic in Michigan began demonstrating conclusively that ageing cannot only be stopped but actually reversed — when elderly patients were obviously rejuvenated and returned to the physical condition of thirty-year-olds — the groundswell of enthusiasm and controversy began to shake the world.

Today, we are quite accustomed to the fact that any man or woman can live as long as he or she desires — barring accident — and remain physically youthful and active indefinitely.

Yet when the first successes of the Fossel Clinic brought that prospect of physical immortality to the full attention of the news media and the general public, great objections were raised and great fears were expressed.

Some feared that only the rich would be able to obtain immortality treatments, and this would lead to a widening and permanent gap between rich and poor. The infamous Death League fomented riots in dozens of cities around the world with their accusations that a wealthy and powerful élite would receive immortality treatments, while the poor and powerless would not be able to afford them.

Misguided though such accusations were, they attracted wide popular support and led to the creation of public committees to oversee the equitable distribution of immortality treatments. Others objected on religious grounds, firm in their belief that attempts to extend the human lifespan were attempts to subvert God's chosen plan. Although the most conservative religious believers firmly rejected immortality treatments for themselves (and, sadly, often for their families), the general public eagerly sought physical immortality.

Theologians then began to ponder the implications of this startling new capability. Many considered, like earlier Church fathers, that any new human faculty can only be attributed to God's beneficence. Some even claimed that Christ's exhortation, "Be

ye therefore perfect, even as your Father which is in heaven is perfect," should be interpreted as referring to physical perfection, as well as moral.

Pope Michael I set the example for many Roman Catholics when he accepted the immortality treatment for himself, while at the same time resigning his position as Pontiff. "No man should remain in such a position of power indefinitely," he wrote in his famous encyclical, *Vitam aeternam*. His moral stance led to widespread public insistence that all political office-holders be subjected to strict limitations on their terms of office.

Pragmatists feared the economic and social dislocations that might accompany immortality, and pointed out that if the death rate were lowered almost to zero, then the birth rate must be equally lowered, or the world would suffer a population explosion of intolerable proportions. Despite these objections, and despite organized political campaigns against immortality (and some terrorist acts, such as the missile attack on the Fossel Clinic), once the general populace understood that death could be postponed indefinitely, an enormous groundswell of public pressure forced governments to abandon all thoughts of suppressing the scientific knowledge or of banning the immortality treatments altogether.

It turned out that the various therapies involved in 'immortalization' were much less expensive than the pessimists predicted. Basically, they consisted of injections of human growth hormones and a cocktail of enzymes such as telomerase. While the research leading to these capabilities was costly, it was largely funded by governmental tax revenues. Because of this, the real costs of

clinical treatments were in fact kept far lower than health economists had expected — more akin to a vaccination than an organ transplant.

The great struggle to adapt society to the greatly extended human lifespan took more than a century, however. Virtually every aspect of social, economic and political life underwent a fundamental change once the dread aspect of death had been thus removed to a distant horizon.

Humanity showed its resilience, in time. Birth rates around the world are very close to death rates now. Global population growth is stabilizing — not without a struggle, but a stable world population is in sight. Today, the birth of a baby is a cause for celebration by the entire community. Couples no longer have children to please their parents, or to gain some perceived advantage in their community. Babies are rare, and cherished.

The advent of immortality has brought an unexpected additional benefit. For all of history, humans have thought the stars were utterly unreachable. Even with modern space technology, it would take centuries to reach the nearest stars. Now, we have centuries to spend. Today the ships bearing the families of scientific explorers are heading out beyond the fringes of the solar system, plying the long, silent pathways to the stars, bearing the seeds of humankind into a waiting Universe.

Ben Bova is the author of more than ninety novels and works of nonfiction. Among the latter is *Immortality: How Science Is Extending Your Lifespan* — and *Changing the World*. He has been Executive Editor of both *Omni* and *Analog Science Fiction* magazine. He is President Emeritus of the National Space Society and a past president of Science Fiction Writers of America.



Survival of the clearest

Steven Pinker

There are no fossils to show how language evolved. But evolutionary game theory is revealing how some of the defining features of human language could have been shaped by natural selection.

The study of the evolution of language, famously banned by the Société de Linguistique de Paris in 1866 and dismissed as idle story-telling ever since, has returned to respectability. Recent years have seen a flurry of articles and books¹⁻⁵, a biannual research conference and now several papers by Martin Nowak and collaborators applying evolutionary game theory to the problem⁶⁻⁹. Their latest offering appears on page 495 of this issue⁹.

Language would seem to be a natural target of evolutionary analysis. Universal, heritable, rapidly acquired by children and well suited to communicating complex information, language is plausibly an adaptation that allowed our ancestors to share knowledge and negotiate agreements, fuelling the technological and demographic explosions that led our species to infest the planet millennia ago.

But the very idea of evolutionary linguistics, like evolutionary psychology in general, draws mixed reactions from scientists. On the one hand, "nothing in biology makes sense except in the light of evolution." The major parts of our psychology could no more be understood without considering their biological function than could the heart or eye. For example, visual perception does not just entertain us with pretty colours and shapes, but delivers usable information about the surfaces and materials of the world. And it is no surprise that people everywhere fear heights and snakes more than comfy chairs and bunnies. On the other hand, other evolutionary 'explanations' of behaviour — "the function of humour is to relieve tension", "music evolved to foster group cohesion" — can seem glib and lame.

Fortunately, good theories of adaptation can be distinguished from bad ones¹⁰. The bad ones try to explain one bit of our psychology (say, humour or music) by appealing to some other, equally mysterious bit (laughing makes you feel better; people like to make music with other people). The good ones use some independently established finding of science or mathematics to show that some mechanism can attain some goal in some environment. For instance, projective geometry shows how stereoscopic imaging can compute the third dimension; herpetology has documented that many snakes are venomous. These engineering bench-



Figure 1 According to recent work in evolutionary game theory — the latest example of which appears elsewhere in this issue⁹ — minimizing communication errors was a major selection pressure in the evolution of human language.

marks can serve as predictions for how Darwinian organisms ought to work. The more uncannily the engineering specifications for a system match the facts of the organism, the more confidently we can infer that the organism was selected to perform that function.

With many behavioural systems, the relevant environment consists not of predictable things like light waves and snakes but of other behaving individuals co-evolving their own strategies. Evolutionary game theory¹¹ has allowed biologists to analyse such dynamics, predicting how Darwinian organisms ought to treat their antagonists, mates and comrades.

Language, like sex, aggression and cooperation, is a game it takes two to play. Many origins have been proposed for it — a substitute for grooming, a way to tattle on adulterers, a by-product of a big head — but they have been based on intuitive appeal. Game theory can provide the external criteria for utility enjoyed by the rest of evolutionary biology. Modellers make the minimal assumption that the transmission of infor-

mation between partners provides them with an advantage (say, by exchanging know-how or coordinating their behaviour), and that the advantage translates into more offspring, with similar communicative skills. The question then is how a stable communication system might evolve from repeated pairwise interactions.

In 1989 the linguist James Hurford¹² used evolutionary game theory to show that a defining property of human language, the arbitrary, bidirectional sign (in which a community of speakers and hearers tacitly agree to use particular signals to refer to particular concepts), will drive out other schemes over evolutionary time. Nowak and collaborators have done the same for two other defining properties of language.

Last year, Nowak and David Krakauer⁶ pointed out that real organisms never transmit information perfectly: errors in signalling or perception are inevitable, especially when signals are physically similar (Fig. 1). Imagine organisms that use a different sound (say, a vowel) for every concept they wish to communicate. As they commu-

nicate more concepts, they will need additional sounds, which will be physically closer and hence harder to discriminate. At some point adding new signals just makes the whole repertoire more confusable and fails to increase its net communicative power.

Nowak and Krakauer showed that this limitation can be overcome by capping the number of signals and stringing them together into sequences, with one sequence per concept. The sequences are what we call words, and the combination of meaningless vowels and consonants into meaningful words (d + o + g = dog) by rules of phonology is a universal property of language: half of what linguists call 'duality of patterning'. Nowak and Krakauer demonstrated that its evolution is likely among communicators with a large number of messages to convey, a precondition that plausibly characterizes our brainy ancestors.

Now, Nowak, Joshua Plotkin and Vincent Jansen⁹ analyse another hallmark. For a word to survive in a community, it must be used frequently enough to be heard and remembered by all the learners. As new words are added to the vocabularies of speakers, old words must be used less often, and they are liable to fade, leaving the language no more expressive than before.

Nowak *et al.* point out that this limitation can be overcome by communicators who use compositional syntax: rather than pairing each word with an entire event, they pair each word with a component of an event (a participant, an action, a relationship), and string the words together in an order that reflects their roles (for example, dog bites man). Such communicators need not memorize a word for every event, reducing the word-learning burden and allowing them to talk about events that lack words. Syntax and semantics, the other half of the duality of patterning, will evolve.

Nowak *et al.*⁹ note that syntax has a cost: the requirement to attend to the order of words. Its benefits exceed the costs only when the number of events worth communicating exceeds a threshold. This 'syntax threshold' is most likely to be crossed when the environment, as perceived by the communicators, has a combinatorial structure, for example, when any of a number of actors (dogs, cats, men, women, children) can engage in any of a number of actions (walking, running, sleeping, biting)¹. In such a world, the number of words that have to be learned by a syntactic communicator equals the sum of the number of actors, actions, places and so on, whereas the number that must be learned by a nonsyntactic communicator equals their product, a potentially unlearnable number. Syntax is handy to an analytical mind in a combinatorial world.

Of course, showing how language could have evolved in a hypothetical world is a far

cry from showing how it did evolve in this world. But the game theorists have demonstrated the evolvability of the most striking features of language — arbitrary signs and duality of patterning — breaking the logical circle that cripples many evolutionary explanations. And by showing how the abstract components of language (lexicon, phonology, syntax and semantics) might be related to more tangible skills, such as the articulation and perception of speech and the cognitive categorization of the world, they suggest ways of connecting the evolution of language to other topics in human evolution, allowing each to constrain the others. Perhaps some day the Société will reverse its ban. ■

Steven Pinker is in the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. e-mail: steve@psyche.mit.edu

Genome sequencing

A view of Mount *Drosophila*

Jonathan Hodgkin

B iologists live in privileged times. In the next few years, the complete genome sequences of representatives from all the major groups of organisms on Earth will be determined. Like a range of mountains emerging from cloud, successively larger and larger genomes can now be seen in their entirety, as each major genome project reaches its goals. Every one of these projects creates a gigantic and breathtaking archive of solid data and new information, which can be analysed in detail for decades to come. Nothing like this has happened in previous scientific history, or is likely to happen again.

The latest grand achievement in genomics has a particular significance for the science of genetics. Almost a century of research on the fruitfly *Drosophila melanogaster* (Fig. 1) has culminated in the publication of 120 million base pairs (Mb) of sequence¹, and it seems to include the entire protein-coding part of the fly genome. Although a number of bacterial (prokaryotic) genomes have been sequenced, this is only the third eukaryotic genome to arrive at this level of coverage. It follows the 12 Mb of the yeast *Saccharomyces cerevisiae*² and the 97 Mb of the nematode worm *Caenorhabditis elegans*³, and is celebrated in a whole series of papers in *Science*^{1,4-9}. These report both the technical aspects of the work and some of the first insights to emerge from a complete panorama of the genome.

Technical issues will not be discussed further here, but are especially important in view of the whole-genome shotgun strategy used for this project, in contrast to the clone-by-clone approach being used by the pub-

1. Pinker, S. & Bloom, P. *Behav. Brain Sci.* **13**, 707–784 (1992).
2. Pinker, S. *The Language Instinct* (HarperCollins, New York, 1994).
3. Hurford, J. R., Studdert-Kennedy, M. & Knight, C. (eds) *Approaches to the Evolution of Language: Social and Cognitive Bases* (Cambridge Univ. Press, New York, 1998).
4. Deacon, T. W. *The Symbolic Species: The Co-evolution of Language and the Brain* (Norton, New York, 1997).
5. Calvin, W. H. & Bickerton, D. *Lingua ex Machina: Reconciling Darwin and Chomsky with the Human Brain* (MIT Press, Cambridge, Massachusetts, 2000).
6. Nowak, M. A. & Krakauer, D. C. *Proc. Natl Acad. Sci. USA* **96**, 8028–8033 (1999).
7. Nowak, M. A., Krakauer, D. C. & Dress, A. *Proc. R. Soc. Lond. B* **266**, 2131–2136 (1999).
8. Nowak, M. A., Plotkin, J. B. & Krakauer, D. C. *J. Theor. Biol.* **200**, 147–162 (1999).
9. Nowak, M. A., Plotkin, J. B. & Jansen, V. A. A. *Nature* **404**, 495–498 (2000).
10. Williams, G. C. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought* (Princeton Univ. Press, 1966).
11. Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, New York, 1982).
12. Hurford, J. R. *Lingua* **77**, 187–222 (1989).



Figure 1 In the limelight — the fruitfly *Drosophila melanogaster*. The sequence of the genome of *Drosophila* has now been published¹, giving geneticists invaluable new information about their favourite organism.

licly funded human genome consortium. It is still unclear whether the shotgun strategy will work when applied to the much larger genomes of vertebrates and economically important plants⁴.

The sequence determined does not correspond to the whole 180 Mb that make up the four chromosomes of the fly, and it still contains gaps and some stretches of low quality. The other 60 Mb (technically, heterochromatin) remains largely uncloned, unsequenced and neglected because most of it consists only of monotonous repeats and dead transposons.

As with most other animal and plant genomes, determining every last nucleotide in the sequence is not justified. The authors declare their sequence "substantially complete", in contrast to the "essentially complete" sequence of *C. elegans*, of which more

than 99% has been determined. The critical question is how much important sequence might still be missing. The problem can be assessed objectively, because partial or complete sequence data are already available for 2,783 fly genes. Only six of these known genes were found to be missing from the assembled sequence, so a reasonable finishing line has been reached.

Now the fun begins for the annotators and the rest of the scientific community. Many broad features of the genome immediately become apparent. One long-standing question, the number of genes in *Drosophila*, has a surprising answer: annotation and gene-prediction programs suggest that there are 13,600, barely twice the number found in yeast and decidedly lower than the estimates for *Caenorhabditis elegans* (more than 19,000, according to current annotation)³ and the flowering plant *Arabidopsis thaliana* (around 27,000)¹⁰, although all three organisms have much the same DNA content (excluding heterochromatin). So the numbers of genes in fly, worm and weed are ranked in exactly the opposite order to that expected from simple estimates of anatomical complexity, cell-type diversity and so on.

However, the differences might be illusory: plants might make less use of alternative splicing of messenger RNA than do animals, instead employing larger gene families. *Caenorhabditis elegans* also carries more duplicated genome segments than *Drosophila melanogaster*, again resulting in more genes. Overall, the effective complexity of the protein complements of the three organisms might be similar. In addition, by some criteria fly proteins are more complex, in that many carry multiple different domains.

In the most fascinating of the articles accompanying the *Drosophila* genome sequence, Rubin *et al.*⁵ compare the contents of the three eukaryotic genomes now available for comprehensive scrutiny. As expected, about 3,000 genes are shared between yeast, fruitfly and nematode, encoding the basic machineries of eukaryotic cell biology. The two animal genomes, however, share many features that are not found in yeast. Rubin *et al.* provide a list of the 200 most frequently occurring protein-coding domains in the fly genome, of which about 25% are completely absent from yeast. In contrast, members of almost all 200 families occur in both fly and nematode, often in remarkably similar numbers.

There are some disparities, but many are explicable: for example, flies have many genes encoding cuticle proteins and chitin-binding proteins, whereas nematodes lack these but have many collagen genes, required for their collagenous cuticles. Some other differences are more a matter of relative usage in major transcription-factor families: flies have more C2H2-class zinc-finger proteins, more homeobox proteins, fewer

nuclear receptors and so on. The general conservation between fruitfly and nematode strongly reinforces the belief that there is a basic gene kit for all animal development and physiology, and one can therefore expect that the distribution of gene families in vertebrates will not be very different.

This expectation is supported by a search for counterparts of human disease genes in the fly genome. Of 289 such genes, 177 (61%) have clear equivalents in the *Drosophila* sequence, several of them being identified for the first time. Examples of new discoveries include genes encoding the tumour suppressor p53 (the 'guardian of the genome'), MEN (involved in multiple endocrine neoplasia), tau and Parkin (both involved in neurological disease). The availability of these as complete sequences in a model organism with so many practical advantages creates a host of new experimental opportunities.

Some properties of the fly genome look both idiosyncratic and intriguing, such as

the large number of trypsin-like proteases (199 in all). No other organism known contains anything resembling this enormously amplified family of proteins. Most striking, however, is the substantial percentage of predicted genes with unknown function and no obvious homologue in any other organism: about 22%. This is a good reminder that even after 90 years of research on the geneticist's favourite animal⁷, there is still a lot to be learned. ■

Jonathan Hodgkin is in the Genetics Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK.
e-mail: jah@bioch.ox.ac.uk

1. Adams, M. D. *et al.* *Science* **287**, 2185–2195 (2000).
2. Goffeau, A. *et al.* *Science* **274**, 546–567 (1996).
3. *C. elegans* Sequencing Consortium *Science* **282**, 2012–2018 (1998).
4. Myers, E. W. *et al.* *Science* **287**, 2196–2204 (2000).
5. Rubin, G. M. *et al.* *Science* **287**, 2204–2215 (2000).
6. Rubin, G. M. *et al.* *Science* **287**, 2222–2224 (2000).
7. Rubin, G. M. & Lewis, E. B. *Science* **287**, 2216–2218 (2000).
8. Kornberg, T. B. & Krasnow, M. *Science* **287**, 2218–2220 (2000).
9. Benos, P. V. *et al.* *Science* **287**, 2220–2222 (2000).
10. Lin, X. *et al.* *Nature* **402**, 761–768 (1999).

X-ray astronomy

Peeking into the obscured Universe

Günther Hasinger

Why is it dark at night? If we lived in an infinite Universe uniformly filled with stars, then every line of sight should eventually end up on the surface of a star and the night sky should appear as bright as the surface of the Sun¹. It was Edgar Allan Poe² who eventually found a solution to this puzzle, known as Olbers' paradox: the stars have not had enough time to fill the Universe with light. Today the brightness of the Universe has been measured over a wide range of electromagnetic frequencies, from radio to high-energy gamma rays (Fig. 1, overleaf), and cosmologists are turning Olbers' question around. What can we learn from the brightness of the sky about the nature and cosmological evolution of objects in our Universe? On page 459 of this issue, Mushotzky *et al.*³ report the first deep observations of the cosmic X-ray background (CXB) with the Chandra satellite — the sharpest X-ray eye around.

The question of whether the CXB is produced by uniform emission from a diffuse intergalactic gas at a temperature of a few million degrees, or by radiation from many individual sources, was settled in the 1990s. First, the COBE satellite did not detect a tiny deviation from the perfect 2.7 K blackbody spectrum of the cosmic microwave background (CMB; Fig. 1), which would be expected if the radiation were seen through a screen of hot intergalactic gas. Second, the majority of the 'soft' X-ray

background (1×10^{17} – 5×10^{17} Hz) was resolved into discrete sources by the X-ray satellite ROSAT. These sources are mostly distant active galaxies (active galactic nuclei, or AGN), containing massive black holes accreting matter from their environment⁴.

The X-ray background should represent the summed emission of all these discrete sources, but the CXB has a spectral shape completely different from the spectra of individual AGN measured at X-ray frequencies. The spectrum of the CXB has a characteristic peak at about 10^{19} Hz (Fig. 1), corresponding roughly to the radiation that medical doctors use to 'X-ray' our bones, whereas the AGN have a flat or slightly concave distribution over the whole X-ray range. A solution to this long-standing puzzle was proposed in 1989, namely that most sources of the X-ray background could be heavily obscured by gas and dust clouds⁵, much like our bones absorb the doctor's X-rays. The higher-energy ('hard') X-rays can penetrate the obscuring clouds, whereas 'soft' photons are absorbed, so that the overall spectrum is different from that of individual AGN.

Models using this idea can reproduce the shape of the observed CXB by assuming sources at various distances and obscured by different amounts of gas⁶. The most important conclusion from these models is that most (80–90%) of the light produced by

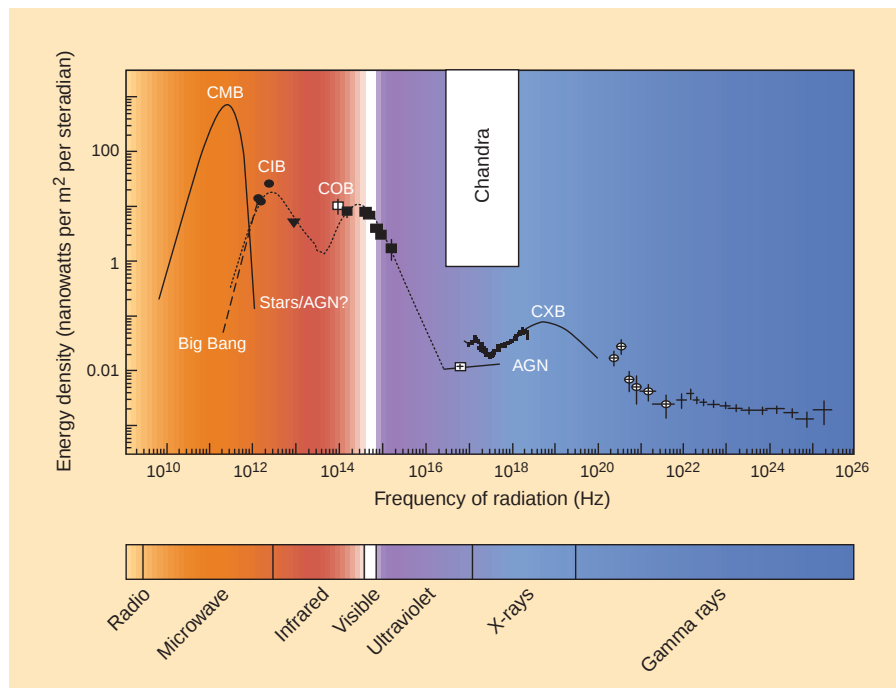


Figure 1 The energy density spectrum of the Universe over the whole electromagnetic spectrum from radio to gamma rays⁷. The four distinct peaks correspond to different cosmologically important emission processes. The 2.7 K cosmic microwave background (CMB) reflects the 'echo of the hot Big Bang'. The recently discovered cosmic infrared (CIB) and optical (COB) backgrounds are probably emitted by warm dust and normal stars, respectively, with an unknown contribution from active galactic nuclei (AGN). The cosmic X-ray background (CXB), discovered 38 years ago, shows a characteristic peak at 10^{19} Hz, which is probably shaped by AGN obscured by dust. The energy band covered by the new Chandra observations³ is shown in white. The vertical axis corresponds to the power emitted by the Universe per energy decade, making it clear that the CMB dominates the total energy output of the Universe.

accretion onto black holes in the early Universe should be obscured, and is probably re-radiated, by warm dust in the cosmic infrared background (CIB; Fig. 1). But because of several uncertain ingredients, the predictive power of these models remains limited.

The new Chandra data expand the range of previous observations significantly. In the soft X-ray band they beautifully confirm and extend the ROSAT results to roughly four-fold fainter limits. In the hard X-ray band (5×10^{17} – 25×10^{17} Hz) they provide the first high-quality images ever. The new Chandra sources must therefore be further away (at higher redshifts) or more obscured than those previously found. The Chandra images resolve at least 75% of the hard X-ray background into discrete sources, and it is reassuring to find that the average spectrum of Chandra sources is now consistent with the background spectrum. The Chandra results broadly agree with models that assume a large population of obscured galactic nuclei, which at optical wavelengths are expected to be very faint or to reside in normal-looking galaxies. This seems to be exactly what Chandra has found.

The most tedious work lies ahead in the optical identification and classification of the X-ray sources and the determination of

their redshift, which is a crucial ingredient for the models. The location of this Chandra survey — the Hawaii deep field — was well chosen for this purpose, because a large number of deep multiwavelength data already exists. But Mushotzky *et al.*³ show that there are many optically faint objects, for which even the giant Keck telescope is not powerful enough to obtain redshifts — the obscured Universe is a hard subject to study. Future observatories such as the Next Generation Space Telescope or the Atacama Large Millimeter Array will be required to study some of the objects that Chandra detected. X-ray spectrophotometry with more powerful X-ray telescopes such as the recently launched Newton (XMM) observatory, or the planned XEUS mission in the distant future, may also help because they might be able to estimate redshifts from X-ray data alone.

One concern in the present study³ is the relatively small size of the chosen survey. The ROSAT experience has shown that a large number of survey fields with different sensitivities and coverages are required for determining the distribution of AGN and their cosmological evolution. In particular the question of whether massive black holes formed at the same time as, or even before, the first galaxies requires much larger sur-



100 YEARS AGO

There are many people who write about experiments upon animals, but only very few who have under their constant notice the actual facts relevant to the subject. In this connection, not merely is a knowledge of fact required, but an intellectuality capable of appreciating the significance of fact. The person most competent from this standpoint is one of the Inspectors under the Act. These Inspectors are most carefully chosen by the Home Office on account of special qualifications which they possess.... It would not, however, be comely for a person holding an official appointment to write a book upon the subject-matter of his office. Every vivisector, a terrible term by which to designate any one who merely pricks a guinea-pig, knows full well that no one, with the above exceptions, is more entitled to write upon the subject of animal experiments than Mr. Stephen Paget, who for twelve years was the active and long-suffering secretary of the Society for the Advancement of Medicine by Research. From *Nature* 29 March 1900.

50 YEARS AGO

For the past three or four years, there has been great congestion of the portion of *Nature* appearing under the title of "Letters to the Editors". Since these 'letters' have come to be regarded as the usual mode of announcement of the results of new work — interim reports in most cases — this was perhaps inevitable in view of the great outburst of scientific activity throughout the world. A further factor was the enforced reduction in size, or even the disappearance, of many of the specialist organs of scientific publication during the war period. One unfortunate result of the great pressure on our space has been the irritating but unavoidable delays in publication, which have been, and are still, a matter of grave concern to the Editors, correspondents and readers. At one period an attempt was made to reduce such delays by using smaller type, so that a greater number of letters could be fitted into the same amount of space. Experience showed, however, that the great inconvenience and strain caused by the use of this small type were not worth the space 'gained'. As soon, therefore, as more paper could be used, we reverted to the usual type and the number of pages devoted to 'letters' was increased.

From *Nature* 1 April 1950.

veys. Fortunately, several groups around the world have started deep Chandra and Newton observations in other, well-studied fields. A great deal of interesting work lies ahead.

Günther Hasinger is at the Astrophysical Institute Potsdam, An der Sternwarte 16, 14482 Potsdam, Germany.

e-mail: ghasinger@aip.de

1. Olbers, H. W. M. in *Astron. Jahrb. für das Jahr 1826* (ed. Bode, J.) 110–121 (Späthen, Berlin, 1923).
2. Poe, E. A. *Eureka, a Prose Poem* (1848).
3. Mushotzky, R., Cowie, L., Barger, A. & Arnaud, K. *Nature* **404**, 459–464 (2000).
4. Schmidt, M. *et al.* *Astron. Astrophys.* **329**, 495–503 (1998).
5. Setti, G. & Wolijer, L. *Astron. Astrophys.* **224**, L21–L24 (1989).
6. Comastri, A., Setti, G., Zamorani, G. & Hasinger, G. *Astron. Astrophys.* **296**, 1–12 (1995).
7. Hasinger, G. in *ISO Surveys of a Dusty Universe* (eds Lemke, D., Stickel, M. & Wilke, K.) (Springer Lecture Notes, Berlin, in the press).

Ecology

Bats about the Arctic

Peter D. Moore

Why do bats fly at night rather than during the day? There are exceptions, of course, but most bats confine their activities to dusk and the depths of night. The insect-eating bats of the higher latitudes, in particular, are night-time hunters and are equipped with sophisticated techniques for the non-visual location of prey. There are several possible advantages that could be associated with nocturnal feeding, such as the avoidance of predators or mobbing by birds, the tapping of a food resource unexploited by other insectivores, or sensitivity to the problems of overheating when active in daylight. In a study of bats, insectivorous birds and predators in northern Norway, Speakman and colleagues¹, reporting in *Oikos*, conclude that the avoidance of competition for food is the most likely answer.

The question has been debated for many years. Nearly ten years ago, Speakman² measured the effects of day-time flying on the activity of bats, and found that they were reluctant to fly in the light, but were not further deterred by high temperatures. The body temperatures of bats after they had flown in the light and in the dark did not differ significantly, so the overheating hypothesis can be rejected. It is also unlikely that bats fly at night to avoid being mobbed by birds, as there are relatively few observations of this occurring and no records of serious harm being done to any bullied bat.

Avoidance of predation or of competition for food therefore remain the most plausible explanations for night flying. Speakman *et al.*¹ have now investigated the relative importance of these two possibilities by choosing a study site within the Arctic circle in summer, for which there is no distinctly dark period and where one might expect to see some corresponding response from birds and bats, together with the higher predators. Their study was based in northern Norway, at latitude 69° N, in early July, when the daylight intensity declines during the period 23:00 to 4:00, but darkness is never total. The temperature also declined during the 'night', from a



Figure 1 The northern bat (top) and nesting sand martins (bottom). It seems that northern bats hunt at night to avoid competing with sand martins for prey¹. This may provide a general explanation for nocturnal flight in bats.

'daytime' high of around 22 °C to a low of about 9 °C between 2:00 and 4:00.

The authors surveyed the abundance of flying insects throughout the day and night, using a simple technique of sweep-netting the air over a defined sample route. By observing a bat roost (an abandoned human dwelling), Speakman *et al.* were able to estimate the activity and flight periods of the northern bat, *Eptesicus nilssonii* (Fig. 1). Similar 24-hour observations of a colony of nesting sand martins (*Riparia riparia*; Fig. 1) provided data about the activity of one of the

important insectivorous birds of the region. The authors also surveyed the hunting activity of avian predators, including raptors such as kestrel (*Falco tinnunculus*), sparrowhawk (*Accipiter nisus*), rough-legged buzzard (*Buteo lagopus*), northern goshawk (*Accipiter gentilis*) and hawk owl (*Surnia ulula*).

Migratory insectivorous birds, such as the sand martin, are able to exploit the abundant insect life of the high-latitude summers and are thought to find the long days advantageous when feeding their young. Speakman *et al.*'s observations showed, however, that there was a period between 24:00 and 3:00 when sand-martin activity ceased and the adult birds were believed to be resting in their nest burrows. Excessive physical stress can suppress the immune system of birds and leave them susceptible to disease³, so a 'nocturnal' rest period seems to be needed. The density of aerial insects corresponded closely with the aerial temperature curve, with lowest insect abundance occurring between 22:00 and 6:00. So, the sand martins choose to rest at a time when insects are less available.

The records of the raptors' activity are less reliable because of their fewer numbers, but overall hunting appeared to be equally intense at all times of the 24-hour day, although different species and individuals were probably involved at different times. The bat activity, on the other hand, was concentrated in the period 22:00 to 2:00. During this time, insect density was at its lowest, but raptor activity continued unabated. Faecal analysis showed that the bats were feeding largely upon small flies of the order Diptera, which are also the preferred food of the sand martins, so there was no indication of insectivores partitioning their resources by hunting different prey, such as moths, during the 'night'.

The evidence indicates, therefore, that bats avoid the feeding time of the birds and are competing for the same food resource. Their pattern of activity does not seem to be related to the risk of predation by predatory birds. But the case is not closed because a few alternative explanations remain. Speakman *et al.*¹ are concerned that their measurements of raptor activity may not correspond with actual risks of predation. As these birds depend on vision for their hunting, the lowering of light intensities in the middle of the 'night' could reduce their efficiency and improve the bats' survival chances. This is not true of all raptors, however. Observations of the kestrel in lower latitudes have shown that this falcon eats bats and can hunt in low-light, crepuscular conditions. It often hunts wood mice (*Apodemus sylvaticus*), a nocturnal mammal, and has even been observed hunting by moonlight⁴.

There is also the problem of the relatively low overall density of both bats and sand martins, which raises the question of

whether the food resource is really limited enough to demand the observed partitioning of feeding times. Competition may not be the immediate cause of the hunting patterns of birds and bats at all. Perhaps the bats in the land of the midnight sun retain a temporal entrainment that has been developed as a consequence of other ecological factors and which, in the short term, is as difficult to escape as human jet lag. ■

Peter D. Moore is in the Division of Life Sciences, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 8WA, UK.
e-mail: peter.moore@kcl.ac.uk

1. Speakman, J. R. *et al.* *Oikos* **88**, 75–86 (2000).
2. Speakman, J. R. *Funct. Ecol.* **5**, 518–524 (1991).
3. Gylfe, A., Bergstrom, S., Lundstrom, J. & Olsen, B. *Nature* **403**, 724–725 (2000).
4. Cramp, S. (ed.) *Handbook of the Birds of Europe, the Middle East and North Africa Vol. II* (Oxford Univ. Press, 1980).

Nuclear physics

Doubly magic nickel

Philip Walker

An atomic nucleus is a collection of neutrons and protons. These come in roughly equal numbers, although a heavy nucleus is likely to choose more neutrons, because they are unaffected by the repulsive electric charges of the protons. But predicting the most extreme combinations of neutrons and protons that will bind together is not so simple. The rules that determine nuclear existence also influence the course of astrophysical processes, such as supernova explosions, that serve to enrich the Universe with heavy elements including those necessary for life on Earth. So, it is an achievement indeed for a team of nuclear physicists to have discovered a new extreme of nuclear matter, as reported by Bertram Blank *et al.*¹ in *Physical Review Letters*. Working at the Grand Accélérateur National d'Ions Lourds laboratory in France, they have created a new isotope of nickel, nickel-48, which is made up of 28 protons and 20 neutrons. This represents the most severe shortage of neutrons for any known isotope.

The reason that 28 protons can support such a paucity of neutrons hinges on the shell structure of nuclei. Just as full electron shells in atoms provide extra stability for the 'inert gases' helium, neon and argon, so full nuclear shells give additional stability to nuclei. Both the proton number (28) and the neutron number (20) of nickel-48 correspond to full shells, called 'magic' numbers by nuclear physicists, so nickel-48 is 'doubly magic'. The new results come after many years of technical development, with a mixture of competition and collaboration between laboratories. The neighbouring isotope nickel-49, for example, was created at the Gesellschaft für Schwerionenforschung laboratory in Germany, with leadership from the same Bertram Blank².

The search for doubly magic nickel-48 began with nickel-58. This is the isotope of nickel with the fewest neutrons that exists naturally, as a stable isotope. A vital feature of the new experiment was making a beam of nickel-58 with high enough intensity. This is

where some chemistry comes in useful. Instead of the usual metallic form of nickel, the volatile compound nickelocene was used to make a plasma, giving a beam intensity of three microamps. Each atom of nickel-58 was accelerated to an energy of four billion electron volts, travelling at about 40% of the speed of light. The beam was directed at a nickel foil target, having the natural mix of nickel isotopes, 68% of which is nickel-58. The idea is that ten neutrons are knocked off a nickel-58 nucleus, but this is such a rare event that complex identification techniques are required.

To pick out the desired nickel-48 events, the atoms that come flying out of the target are sent through a series of magnets and particle detectors, which measure such things as position, time and energy loss. Together with other reaction products — about 500 of them per second — the nickel-48 atoms are finally stopped by a stack of five silicon detectors, again recording position, time and

energy. Each stopped atom produces ten independent parameters that could be used to pin down its identity. That sounds like overkill, but in the end Blank *et al.*¹ were able to identify just four nickel-48 events (Fig. 1), with the qualification that it is "rather unlikely that more than one of these four events is background". After ten days of bombarding the nickel target with three microamps of nickel-58 beam (about a million million nickel-58 atoms per second) the confident identification of three or four nickel-48 atoms requires precise evaluation.

Was it worth it? The answer is an emphatic 'yes'. We now have a window into a variety of fascinating properties that nickel-48 might exhibit. For example, nickel-48 is expected to decay rapidly by two-proton radioactivity, itself an unexplored process. Further, the new results suggest that the half-life of nickel-48 must be at least a microsecond — long enough for the atoms to fly through the experimental apparatus and reach the silicon detectors at the end. But a microsecond is already a much longer half-life than predicted by some theories. So the data are already specifying the necessary ingredients to describe such a neutron-deficient nucleus, and improved techniques are expected to lead to more-detailed results in the future.

Nickel is unique in having three doubly magic isotopes. Nickel-56 is doubly magic; and, with 50 neutrons, doubly magic nickel-78 is actually one of the most neutron-rich nuclei. And there is another twist to the latest discovery. If the proton and neutron numbers are exchanged, the nucleus obtained is called a 'mirror' nucleus. Because the nuclear force is independent of the neutron–proton swap, any differences in properties arise almost entirely from the different charge

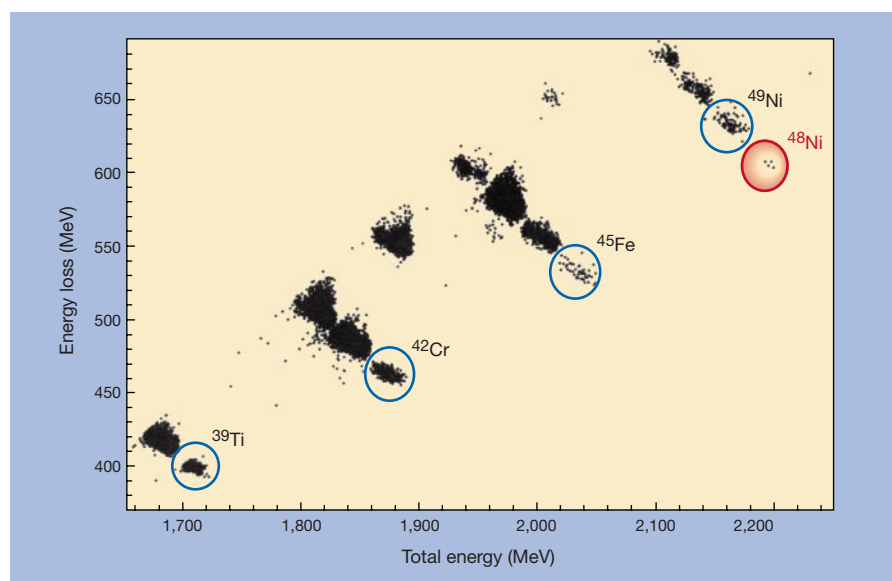


Figure 1 Identification of four nickel-48 atoms by Blank *et al.*¹. The plot shows energy loss in the first silicon detector against the total energy from the sum of the first three silicon detectors. The nickel-48 events are circled, along with other exotic isotopes (from ref. 1).

interactions: protons are positively charged, neutrons are neutral. This simplifying feature has long been a valued source of information about the way in which neutrons and protons bind together. For nickel-48, its mirror nucleus is calcium-48, which is stable, and both are doubly magic. This is the only known pair of doubly magic, mirror nuclei, and no other such pair is ever likely to be

found. So, in the future, identifying excited states of nickel-48 should allow for interesting comparisons with the excited states of calcium-48.

Philip Walker is in the Department of Physics, University of Surrey, Guildford GU2 7XH, UK. e-mail: p.walker@surrey.ac.uk

1. Blank, B. *et al.* *Phys. Rev. Lett.* **84**, 1116–1119 (2000).

2. Blank, B. *et al.* *Phys. Rev. Lett.* **77**, 2893–2896 (1996).

Enzymology

Not just an active site

David W. Banner

Enzymes are biological catalysts. At their simplest, they are proteins that bind a substrate molecule in a special pocket, known as the active site, on their surface. They then convert the substrate into product by enabling a chemical reaction to take place more rapidly than it would in free solution. Enzymes are often involved in the control of complex processes in living organisms, so their functions must be carefully regulated. This frequently occurs by the use of so-called 'exosites' on the enzyme surface that are distant from the active site. On pages 518 and 465 of this issue, Fuentes-Prior and col-

leagues¹ and Dennis and co-workers² offer fascinating new insights into the structure, function and medical importance of exosites in two enzymes involved in the blood-clotting process.

Blood is a marvellous fluid that not only transports food and oxygen to tissues and removes waste products, amongst many other functions, but also has the ability to repair damage to blood-vessel walls by blocking the holes with blood clots — a process known as haemostasis. Clots that grow too large or break away from their site of origin cause the blockage of blood vessels,

or thrombosis. The balance between haemostasis and thrombosis is maintained by controlling the level of an enzyme known as thrombin³.

The enzymes involved in haemostasis and thrombosis are all initially produced in an inactive 'zymogen' form and are activated by cleavage of specific peptide bonds. All of the enzymes and their cell-surface co-factors are modular, consisting of several different structural domains. This type of architecture (Fig. 1) is normal for cell-surface proteins and introduces the possibility, in this case, of controlling enzyme function by means of interactions between two or more proteins, or between proteins and regulator molecules, such as calcium or heparin.

When the wall of a blood vessel is damaged, circulating factor VII, a small fraction of which is already 'active' (denoted factor VIIa), forms a complex with its co-factor, cell-bound tissue factor. This so-called 'extrinsic Xase' complex starts the coagulation process (Fig. 1). The crystal structure of this complex⁴ shows how the rigid tissue factor localizes and stiffens the floppy factor VIIa. The complex then binds factor X, and the active site of factor VIIa can cleave the peptide bonds required to generate activated factor X (factor Xa) from factor X. Factor Xa may also be generated by a tissue-factor-independent pathway, which maintains pro-coagulant activity through 'intrinsic Xase'. This complex contains factor IXa and factor VIIIa, and is structurally similar to the complex between factors Xa and factor Va.

Factor Xa and factor Va form the 'pro-thrombinase' complex, which releases thrombin from prothrombin. A feedback loop ensures that there is an explosive local release of thrombin. Thrombin then converts fibrinogen to fibrin, a long rod-like molecule, which self-associates to make a network that is the basis of a blood clot.

It is then essential to shut coagulation down in a regulated fashion. Central to this process is a molecule called thrombomodulin (Fig. 1), which binds thrombin and turns it into an anticoagulant^{5,6}. How does this transformation in function occur? Insight is provided by Fuentes-Prior *et al.*¹, who present the structure of thrombin bound to the central part of thrombomodulin (epidermal growth factor (EGF)-like domains 4, 5 and 6).

The thrombomodulin–thrombin complex shuts down coagulation by activating protein C, which complexes with protein S and inactivates factors Va and VIIIa by degradation. Before the study of Fuentes-Prior *et al.*, it was unclear whether or not the initial activation of protein C occurred as a result of a change in the active site of thrombin. Such an alteration might be either direct or 'allosteric' — that is, the result of a conformational change caused by binding of thrombomodulin elsewhere on the enzyme.

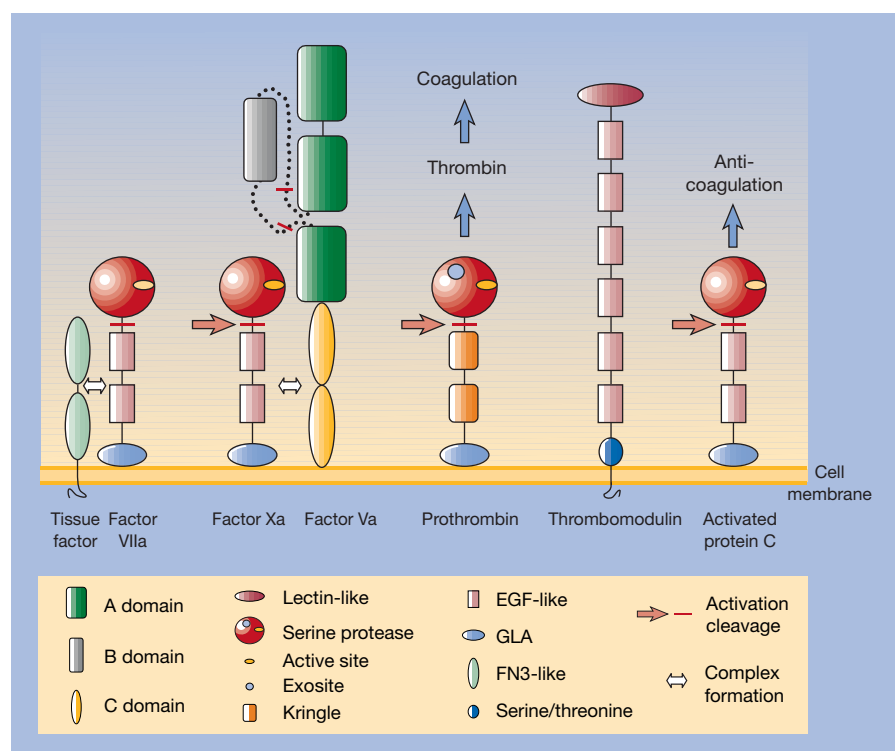


Figure 1 Some of the molecules involved in blood coagulation. Extracellular protein domains are shown as filled shapes. EGF, epidermal growth factor; GLA, γ -carboxy-glutamic-acid-rich; FN3, fibronectin-3. Starting from the left, tissue factor and factor VIIa come together; factor X is then activated (denoted factor Xa), binds to factor Va and, in turn, activates prothrombin to thrombin. Thrombin initiates coagulation by cleaving fibrinogen to fibrin (not shown). Once bound to thrombomodulin, thrombin switches from procoagulant to anticoagulant activity, acting through protein C.

The structure obtained by Fuentes-Prior *et al.* shows thrombomodulin binding in the well-studied anion-binding exosite of thrombin⁷, which is where thrombin binds fibrinogen. So, thrombomodulin both inhibits fibrinogen binding and provides a protein-C-binding exosite, well away from the thrombin surface, that is positioned in such a way that the bound protein C will have its 'activation peptide' placed correctly as a substrate in thrombin's active site.

The representation (Fig. 1) of the domains of modular proteins such as thrombomodulin as a 'string of beads' is useful for many purposes⁸, but begs the question of whether the protein is rigid or floppy. Fuentes-Prior *et al.*'s three-dimensional structure of a thrombomodulin fragment in complex with thrombin shows the three EGF-like domains forming a well-ordered and unexpected structure. In this and many other cases, the protein domains assemble together into a compact, coherent, three-dimensional arrangement that is essential for their biological function.

In the second paper, Dennis *et al.*² take an intriguing approach to the question of how to identify exosites and use them to develop medicines, in this case anticoagulants. The authors expressed libraries of cyclic peptides on the surface of bacterial viruses (bacteriophages), and selected out, in an iterative way, those peptides that show the highest affinity for factor VIIa. They discovered peptides that do not interfere with formation of the complex between tissue factor and factor VIIa, and that bind tightly to a site on factor VIIa that is distinct from the active site. They have thus identified a previously unknown

exosite, and have made the remarkable discovery that the best 'phage peptide' binds specifically to this site on factor VIIa and totally blocks the processing of factor X.

Although medicinal chemists have made much progress in producing inhibitors of enzyme active sites (see ref. 9 for a review), there have been few advances in targeting protein-protein interactions (the development of hirugen being one exception¹⁰). As the enzymes involved in coagulation, as well as many other biologically important enzymes, are closely related members of the thrombin/trypsin family, there is also a problem of selectivity when attempting to inhibit just one particular enzyme, with the risk of producing the wrong drug profile or toxic side effects. Dennis *et al.*² have offered a way of circumventing these problems. Time will tell whether molecules of this sort, or their chemical mimetics, will be the highly specific anticoagulant drugs of the future. In the meantime, we would do well to remember the importance of exosites, and that an enzyme is never just an active site. ■

David W. Banner is in the Department of Chemical Technologies, Pharmaceutical Research Division, F. Hoffmann-La Roche, CH 4070 Basle, Switzerland.
e-mail: david.banner@roche.com

1. Fuentes-Prior, P. *et al.* *Nature* **404**, 518–525 (2000).
2. Dennis, M. S. *et al.* *Nature* **404**, 465–470 (2000).
3. Griffin, J. H. *Nature* **378**, 337–338 (1995).
4. Banner, D. W. *et al.* *Nature* **380**, 41–46 (1996).
5. Esmon, C. T. & Fukudome, K. *Semin. Cell Biol.* **6**, 259–268 (1995).
6. Esmon, C. T. & Mather, T. *Nature Struct. Biol.* **5**, 933–937 (1998).
7. Berliner, L. J. *et al.* *Biochemistry* **24**, 7005–7009 (1985).
8. <http://smart.embl-heidelberg.de>
9. Leung, D., Abbenante, G. & Fairlie, D. P. *J. Med. Chem.* **43**, 305–341 (2000).
10. Naski, M. C. *et al.* *J. Biol. Chem.* **265**, 13484–13489 (1990).

mixed with quantum mechanics, as shown by Blaauwgeers *et al.* for the special case of rotating superfluid ³He (the lightest isotope of helium).

Uniquely, helium at low pressure remains liquid down to absolute zero. The laws of thermodynamics require all materials to be fully ordered at absolute zero, and because crystalline order is not possible, liquid helium must achieve perfect order by another route. As zero temperature is approached, the atoms in the liquid fall into the lowest available quantum-mechanical energy state, in which individual atoms lose their identities and become components of something more akin to one giant liquid molecule filling the whole volume. Because collisions between atoms are ruled out if they all move in unison, there can be no friction. These liquids are therefore 'superfluids' and under the right conditions can be set into motion, which will persist for eternity.

Vortices in such superfluids are very odd. As all the atoms are in the same quantum state they can be described by a single wavefunction — the quantum wave pattern that contains all knowable information on the state of the liquid. Introduce a defect, such as a vortex, and the wave pattern must wrap itself around the core and join up with itself seamlessly. This means that only certain types of vortex structure are possible, because the number of wavefunction oscillations around the vortex core must be a whole number.

The number of these oscillations also defines the 'circulation' of the fluid, which is quantized in steps given by 1, 2, 3, ..., *n* units of 2 π , where 2 π corresponds to one wavefunction oscillation around the vortex. Whereas the angular momentum of the liquid in the vortex increases linearly with this 'winding number', *n*, the kinetic energy of the motion increases as the square. Because a multiply quantized vortex has much more kinetic energy than an equivalent number of singly quantized vortices, the latter are preferred on quantum grounds. Until now only singly quantized vortices have been identified.

The rare isotope ³He is even more unusual. The ³He atom is a fermion and therefore obeys Pauli's exclusion principle, which allows only one particle per state. Hence, for ³He atoms to fall into the lowest state, they must do so in pairs — the so-called Cooper pairs. (Two fermions make a boson, which is not restricted by Pauli's law.) Because the Cooper pairs in ³He have an orbital angular momentum of one unit and a nuclear spin of one unit, the wavefunction of this superfluid must tell us the direction in which the orbital and spin momenta are pointing as well as the liquid velocity. This extra complexity yields a whole hierarchy of possible superfluid phases with different geometries. Moreover,

Superfluidity

A new twist to an old story

George Pickett

Elsewhere in this issue (*Nature* **404**, 471–473; 2000), R. Blaauwgeers *et al.* describe the identification of a new type of quantum vortex, which appears when superfluid helium-3 is rotated.

There are two ways to rotate a liquid. Spin a bucketful and the whole lot will turn as one unit, with each particle in the liquid maintaining its relative position with respect to the bucket. In contrast, the angular momentum of the particles varies with position and there are violent consequences if you try stirring at the same time (as violent storms in the tropics demonstrate). Further, because the velocity of the liquid increases with distance from the centre there must be some outer boundary (in this case the bucket), otherwise the rotating outer layers will eventually exceed the velocity of light.

The other way to rotate the liquid is to let water in the stationary bucket drain through a hole in the base. The escaping water forms a vortex, just as it would down a plughole. In a vortex the velocity decreases inversely with distance from the axis and thus there must be some inner boundary towards the centre, otherwise the velocity will approach infinity close to the axis. Water draining through a plughole solves this problem by removing itself from the axis to leave an empty core. In a vortex, the angular momentum is the same everywhere and vortex structures are consequently very stable.

Vortices are common in nature and occur on all scales, from cosmic strings (a vortex in the structure of space-time) down to atomic dimensions. Some of the most interesting vortices appear when vortex behaviour is

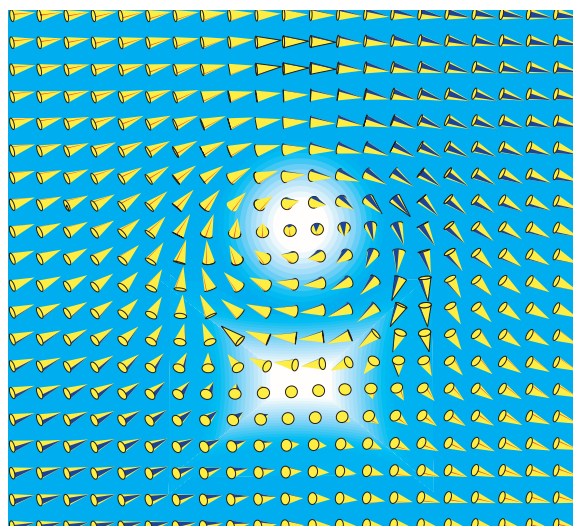


Figure 1 The doubly quantized vortex observed by Blaauwgeers and colleagues. The cones indicate the direction of the Cooper-pair angular momentum. The angular momentum makes a full turn in direction on going round the circular and hyperbolic cores. The circulation is given by the rotation of the cones about their axes. This makes a full 4π , double turn, following round the perimeter.

because the direction of both the nuclear and the orbital spins, as well as the phase of the wavefunction, can make a 2π change around the axis, then many different vortex structures are possible and the energy restriction on all but singly quantized vortices is lifted.

In their experiment, Blaauwgeers and colleagues identify a doubly quantized vortex (that is, $n=2$) in superfluid ^3He from its NMR signature. This vortex is shown in Fig. 1. The cones point in the direction of the orbital angular momentum of the Cooper pairs. The circulation depends on the angle of rotation of the coloured bands around the cone axis. Following around the perimeter of Fig. 1, the cones make two full turns, corresponding to a 4π rotation.

In ^3He , several quantities are changing simultaneously, so there is no singularity along the axis (the air-filled core in the plug-hole version) and all quantities vary smoothly across the axis. The vortex is also rather large, of order 10 micrometres in diameter, or not far off the size of a small snowflake. Indeed, this is a new snowflake, a natural

structure having its own beautiful symmetry and the added bonus of stable symmetrical flow. Because superfluid ^3He can only exist below 2.6 mK above absolute zero, these new 'snowflakes' are even more fugitive than those of ice.

Vortices in this superfluid are not simply aesthetic structures in an arcane fluid. The symmetry of the wavefunction of superfluid ^3He mimics surprisingly closely the metric of the Universe. Vortices are the superfluid analogue of cosmic strings. In the Universe, cosmic strings, if they indeed exist, are the relics of cosmic turbulence frozen into space-time by rapid phase transitions in the Universe shortly after the big bang. In the experiment of Blaauwgeers *et al.* the vortices are under strict control and can be injected one by one into the liquid as the rotation rate is increased. Indeed, this is how the experiment is done. Until we find real cosmic strings to study, vortices in superfluid ^3He are the best substitute.

George Pickett is in the Department of Physics, University of Lancaster, Lancaster LA1 4YB, UK. e-mail: g.pickett@lancaster.ac.uk

Bacterial genomics

ABC of meningococcal diversity

Ian M. Feavers

Meningococcal disease is dreaded by parents and medical practitioners alike for its rapid onset and the difficulty of obtaining a timely and accurate diagnosis. Although the causative bacterium, *Neisseria meningitidis* (Fig. 1, overleaf), is responsive to antibiotic therapy, these features of the disease pose a particular problem for treatment. Prevention of the disease through effective immunization is thus a more attractive alternative than therapy. For example, during the past five months new

vaccines have been introduced into the UK vaccination programme that, for the first time, offer protection to young infants against disease caused by *N. meningitidis* serogroup C. The success of this programme emphasizes the need for the development of vaccines against all pathogenic strains of *N. meningitidis*.

The complete genome sequences of the serogroup A and B organisms are now reported by, respectively, Parkhill *et al.* on page 502 of this issue¹ and Tettelin *et al.* in

*Science*². These papers represent a milestone in meningococcal research. As well as providing the basis for future studies of the biology and the disease-causing mechanisms of this important pathogen, they highlight the potential of bacterial genomics to provide mankind with new solutions to intractable infectious diseases. However, the data also show us that a headlong rush to evaluate new vaccine candidates should be tempered with caution. Both genomes are replete with an unprecedented number of genetic elements that contribute to both genome fluidity and physical variability, attesting to the adaptability of *N. meningitidis* and its potential to evade the human immune response.

Neisseria meningitidis is a major cause of bacterial meningitis and septicaemia, diseases that all too frequently result in death or disability. Serogroup A organisms are responsible for large epidemics in sub-Saharan Africa, whereas serogroup B and C organisms account for sporadic cases and numerous localized outbreaks worldwide³.

Despite the reputation it has earned as an aggressive pathogen, *N. meningitidis* is more frequently found as a harmless commensal — an organism that lives with, but does not harm, the host — in the human nose and pharynx. It has been estimated that, at any given time, between 5% and 10% of people carry the organism. Sporadic cases of disease occur at the much lower annual rates of 1 to 5 per 100,000. The bacterium lives only in humans, passing from one person to another in aerosols; there are no other animal or environmental niches that it can colonize. It is therefore highly adapted to living in humans and has evolved a variety of mechanisms to evade our defences, including the ability to produce polysaccharide capsules that provoke a poor immune response, and very variable cell-surface proteins that represent an ever-changing target for the immune system.

The genetic basis for the adaptability of *N. meningitidis* is evident from the analyses of the genome sequence data by Parkhill *et al.*¹ and Tettelin *et al.*². These data reveal an abundance of diverse repeat sequences in each meningococcal strain. These sequences are likely to have a key role in mediating variation within each genome and hence in promoting the adaptability of the organism. A large proportion of repeat sequences is associated with genes encoding cell-surface structures that are involved in the interaction of the organism with the host and thus represent possible targets for vaccine development. Tettelin *et al.*² estimate that the expression of as many as 65 genes in the serogroup B strain MC58 might be altered by the gain or loss of nucleotides within iterative sequence motifs as a result of inaccurate DNA replication. The effect of such sequence changes on the expression of these genes and, ultimately, on the synthesis of the proteins they encode,

provides the bacterium with the ability to switch rapidly and reversibly between a vast number of possible combinations of physical attributes.

Other types of repetitive DNA allow further diversity, mediating genomic recombination both within and between different meningococcal strains. Recombination between strains makes use of the natural ability of the meningococcus to acquire DNA from its environment and is undoubtedly helped by the prevalence of DNA-uptake sequences — specific nucleotide sequences that enable the acquisition of DNA by other meningococci — in the genomes. All of these mechanisms provide the bacterium with the means to overcome the host's immune response to both infection and vaccination. Such concerns have been raised recently with regard particularly to the ability of meningococcal strains to exchange the genes responsible for biosynthesis of their capsules⁴.

Fortunately, as well as highlighting the extent of the problem, meningococcal genomics also provides possible solutions. Pizza *et al.*⁵ have screened the serotype B genome for genes encoding cell-surface antigens that are highly conserved in all meningococcal strains. Their aim is to make use of these antigens as vaccine components. This approach, although giving encouraging results in an animal model⁵, must now be proved in clinical studies. The comparison of different meningococcal genomes might also reveal genes that distinguish the relatively small number of hypervirulent strains from all other meningococci, and provide the basis for a strategy that specifically targets the properties encoded by such genes.

We do not yet know which strategy, or combination of strategies, will provide an

effective solution to the problem of meningococcal disease. But the public availability of the two genome sequences is bound to hasten the development of comprehensive vaccines that offer protection against all virulent meningococci.

Ian M. Feavers is in the Division of Bacteriology, National Institute for Biological Standards and

Control, Blanche Lane, South Mimms, Potters Bar EN6 3QG, UK.

e-mail: ifeavers@nibsc.ac.uk

1. Parkhill, J. *et al.* *Nature* **404**, 502–506 (2000).
2. Tettelin, H. *et al.* *Science* **287**, 1809–1815 (2000).
3. Schwartz, B., Moore, P. S. & Broome, C. V. *Clin. Microbiol. Rev.* **2** (Suppl.), 118–124 (1989).
4. Maiden, M. C. & Spratt, B. G. *Lancet* **354**, 615–616 (1999).
5. Pizza, M. *et al.* *Science* **287**, 1816–1820 (2000).

Particle physics

Backyard exotica

Frank Wilczek

All atoms follow the same body plan. They consist of a small, massive nucleus of positive charge surrounded by a cloud of negative electrons, bound together by the electromagnetic force. One might say that all atoms belong to a single phylum. Hadrons, elementary particles that are governed by the strong nuclear force, are known to populate three extremely well documented phyla. Baryons (such as protons and neutrons) are based on three quarks having different colour charges, one each of red, white and blue. Antibaryons are similarly constructed from triads of antiquarks, and mesons are based on bound quark–antiquark pairs, carrying equal and opposite colour charges. Each of these hadron phyla supports dozens of species. For many years now, physicists have been on the lookout for exotic new species, requiring more hadron phyla. We might now have the first hint from a theoretical analysis by Alford and Jaffe¹ of something suitably exotic.

What are these new particles? Among the range of possibilities dreamt up by theorists are various types of glueball, made from either two or three colour gluons (the particles that bind quarks together). Or there are centaurs (also called hybrids) made from a quark, an antiquark and a gluon. Finally, we have a hypothetical phylum that does not yet have an accepted name, based on two quarks and two antiquarks. It seems irresistible to call these quarklets. Alford and Jaffe use ingenious computer experiments to argue that some long familiar, but hitherto enigmatic, particles are best interpreted as quarklets¹.

Hundreds of different strongly interacting particles have been identified. Most are very short-lived (less than 10^{-20} seconds), and are observed only as enhancements in particle production rates, known as resonances. The observed properties of many of these particles match up with the predicted properties of different arrangements of quarks and antiquarks conforming to one of the baryon, antibaryon or meson body plans. Within each plan, one has the freedom to choose different properties for the quarks,

which can result in states with different total energies (or masses, according to $E=mc^2$). By this reasoning, the pattern of resonances can be seen as a 'spectrum', wholly analogous to the pattern of stationary states of an atom (as observed from its spectral lines).

The logical development of this line of thought is the quark model². The quark model was cultivated during the 1960s and early 1970s, finally reaching its most mature form in the so-called bag model³. According to the bag model, hadrons consist of relativistic, nearly free quarks (and/or antiquarks) confined to a finite region of space by a universal, constant vacuum pressure. The quark/bag model accounts, semi-quantitatively, for an enormous range of data. But it is not a satisfactory fundamental theory. For example, simply trying to compute what happens when bag-hadrons move results in mathematical ambiguities.

The fundamental theory of the strong force, quantum chromodynamics or QCD, emerged from the study of extremely high-energy collisions. QCD accepts quarks and antiquarks as fundamental ingredients, but also requires colour gluons. Indeed, the exchange of colour gluons is fully responsible for the strong interactions between quarks, much as the exchange of photons is responsible for electromagnetic interactions. Unlike the quark/bag model, QCD is a proper, precisely defined fundamental theory.

If correct, QCD should be able to predict the major properties and strong interactions of hadrons. Unfortunately, however, the equations of QCD are very difficult to solve. Computations that strain at the limits of modern, massively parallel computers succeed in producing, within a few per cent accuracy, the masses of a dozen or so of the lowest states of the hadron spectrum⁴. This is a remarkable achievement and confirms the correctness of a beautiful, parameter-free theory⁵. Yet to understand the rest of the hadron spectrum one will need, for the foreseeable future, to build upon older, but more manageable, concepts and models. Within this humbler approach, QCD does not dic-

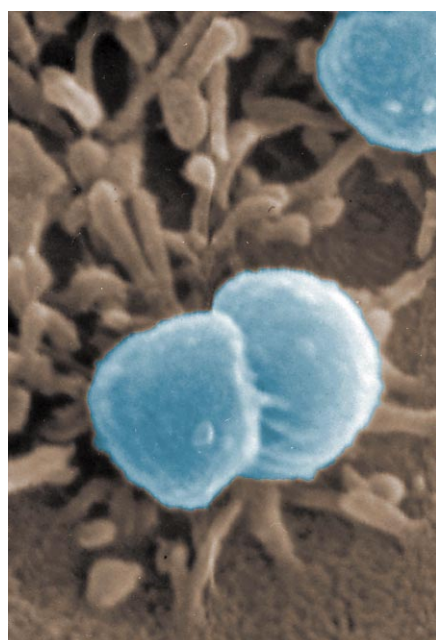


Figure 1 The meningococcus — a bacterium with unprecedented potential for diversity^{1,2}.

tate answers, but rather suggests ideas. For example, the overall colour-charge neutrality of three-quark baryons and quark-antiquark mesons explains why they have special status as stable objects. So QCD sanctions the semi-empirical approach of the quark/bag model, much as quantum theory validates the chemists' models of valence and bonding.

But it is important, and perhaps more exciting, that beyond endorsing old working assumptions, QCD suggests new questions and possibilities. Notably, QCD predicts each of the alternative body plans described earlier. Such beasts ought to exist, and they probably do. Unfortunately, particles or resonances observed in the laboratory do not come tidily labelled 'glueball', 'centauron' or 'quarktet'. Although candidate glueballs and centaurons have been vigorously advocated, other more conventional interpretations have not yet been firmly excluded.

The new work¹ concerns a family of resonances known as 0^{++} mesons, with names such as $\sigma(660)$, $f_0(980)$ and $a_0(980)$. 0^{++} refers to the spin (0), parity (even or +) and charge conjugation parity (+). (I won't try to explain these technicalities here.) The numbers in parentheses are the approximate masses of the resonances in megaelectron volts. (In these units, the electron mass is 0.5 and the proton mass is 938.) Here, a_0 actually refers to a related trio of particles, a so-called isotriplet.

If we attempt to explain the f_0 and a_0 particles within the conventional meson body plan, then we must combine either an up or a down quark with either an up or a down antiquark, using special, rather complicated wavefunctions. But then one would expect to find another heavier meson of the same kind, made from the heavier strange quark and its antiquark, whereas the σ is found to be lighter. On the other hand, the quarktet picture provides a very pretty, though slightly intricate, explanation of the observed pattern, based on the fundamental forces between quarks^{6,7}.

Such arguments⁸ are already quite suggestive, but Alford and Jaffe have raised the discussion to a new level, by formulating and performing a crucial (numerical) experiment. In pure QCD, the 0^{++} particles decay rather rapidly into pairs of 'pseudoscalar' mesons (π , K , η), which complicates their study and interpretation. So Alford and Jaffe study a slightly modified version of QCD, in which the masses of the quarks are taken to be larger than their real-world values, where this can't happen. They also 'improve' on reality by removing the possibility of quark-antiquark annihilation, thereby sharpening the distinction between mesons and quarkkets. They find that pairs of pseudoscalar mesons feel attractive forces, which will cause them to form stable bound states

whose qualitative properties (spin, parity and mass) match those of the observed 0^{++} particles. As each pseudoscalar is a quark-antiquark meson, the combination is a quarktet!

More accurate and extensive work is required to prove that the bound states really do form, and to verify that nothing untoward happens when we swap modified for real QCD. If everything checks out as expected, then in the search for exotic hadrons it seems that, after voyages into realms of extravagant energy and strangeness, we "arrive where we started, and know the place for the first time"⁹.

Frank Wilczek is at the Institute for Advanced Study, School of Natural Sciences, Olden Lane, Princeton, New Jersey 08540, USA.

e-mail: wilczek@sns.ias.edu

1. Alford, M. & Jaffe, R. <http://xxx.lanl.gov/abs/hep-lat/0001023>
2. Kokkedee, J. *The Quark Model* (W. Benjamin, New York, 1968).
3. DeGrand, T., Jaffe, R., Johnson, K. & Kiskis, J. *Phys. Rev. D* **12**, 2060–2076 (1975).
4. Wittig, D. <http://xxx.lanl.gov/abs/hep-ph/9911400>
5. Wilczek, F. *Nature* **397**, 303–306 (1999).
6. de Rujula, A., Georgi, H. & Glashow, S. *Phys. Rev. D* **15**, 281–289 (1977).
7. Jaffe, R. L. <http://xxx.lanl.gov/abs/hep-ph/0001123>
8. Black, D., Fariborz, A., Sannino, F. & Schechter, J. *Phys. Rev. D* **59**, 074026–074038 (1999).
9. Eliot, T. S. from *Little Gidding in The Four Quartets* (Harcourt Brace, San Diego, 1974).

Ancient DNA

Neanderthal population genetics

Matthias Höss

Authenticity is all in research on ancient DNA. Experience has taught us that even the most exciting claims of the retrieval of ancient DNA are not worth much if they cannot be independently reproduced. Hence the importance of a paper on page 490 of this issue, in which Ovchinnikov *et al.*¹ describe the extraction, amplification and sequencing of DNA from 29,000-year-old archaeological bone material of a Neanderthal recovered from the Mezmaiskaya Cave in the northern Caucasus (see map on

page 490). This is the second time that such a claim has been made, the first being in 1997 (ref. 2). The paper by Ovchinnikov *et al.* is probably the more important of the two, for it provides invaluable corroboration for the authenticity of Neanderthal DNA sequences. Moreover, sequences of the DNA from a second Neanderthal offer more detailed insight into the contentious evolutionary relationship between Neanderthals and modern humans.

Research into ancient DNA enjoys high

Box 1 Studying ancient DNA

Research on ancient DNA developed in the mid-1980s from molecular evolutionary research, with the aim of extending phylogenetics (studies of evolutionary relationships) and population genetics to extinct species and populations.

Mitochondrial DNA (mtDNA) is used mainly for research into ancient DNA. Such DNA has advantages for evolutionary research in general. In addition, the large number of copies of mitochondrial genes (up to 1,000 times more copies of mitochondrial DNA are available per cell, compared with single-copy nuclear genes) makes mtDNA a better option when hunting for surviving ancient DNA molecules.

The polymerase chain reaction (PCR) is the most important tool in this field. The

exponential DNA amplification provided by this technique is highly efficient and, in theory, a single DNA molecule can be amplified to provide enough material for sequencing. Without PCR, it would be impossible to study the sequences of ancient DNA molecules.

However, contamination of samples by modern DNA molecules is a severe problem, in part because of the efficiency of PCR. Because only tiny amounts of target (ancient) DNA molecules tend to be retrieved, even low levels of contamination with modern DNA will distort the result. Human remains are particularly difficult to work with, owing to the difficulty of identifying contaminating, generally human, sequences.

Water and oxygen, which

cause hydrolysis and oxidation of DNA, respectively, are the main causes of DNA decay over time. Any burial conditions that limit or exclude water and/or oxygen will improve the survival of DNA in archaeological samples. Low temperatures during the burial period also seem to increase the chance of DNA survival, by slowing its breakdown by hydrolysis and oxidation. The age limit for successful retrieval of DNA depends to a large extent on the burial conditions. So far, the upper limit seems to be around 100,000 years.

Finally, the resurrection of extinct species is, and will stay, impossible. The reason? Only small, insufficient amounts of ancient DNA can ever be recovered, even under ideal conditions.

M.H.

publicity. It is perhaps the combination of modern molecular techniques and 'old-fashioned' archaeology that catches the interest of the scientific community and general public alike. This fascination sometimes clouds critical judgement. But this area of research, like all others, must meet with standards that ensure the authenticity and reproducibility of any given result. This has not always been so. Several of the most spectacular claims — such as the retrieval of DNA sequences from 15-million-year-old plant compression fossils³, from 80-million-year-old bones of putative dinosaur origin⁴ and from insects of up to 130 million years in age trapped in amber^{5–7} — could not be reproduced in any other than the original laboratories, and so are of limited value^{8–11}.

The relationship between Neanderthals and humans remains enigmatic, so the retrieval of Neanderthal DNA has been one of the major goals of researchers in the field of ancient DNA. The age of later Neanderthal populations is well within the range compatible with reliable retrieval of ancient DNA (such retrieval is possible from samples up to 100,000 years old). However, it appeared from several studies (for example, ref. 12) that the work done with ancient human remains was close to the technological limit of what is possible. This is mainly because of the difficulty of distinguishing target sequences from contaminating modern, in this case human, DNA (Box 1).

So it came as no surprise that the publication of the first successful retrieval of DNA from a Neanderthal, from the Feldhofer Cave in Germany², was greeted with caution. Although the paper was widely regarded as being of technically high quality, the remote possibility remained that the published sequence was an artefact or the result of contamination. The need for DNA sequences from a second, unrelated Neanderthal specimen was clear, as echoed in most reviews of that paper. And this is where the importance of the work of Ovchinnikov *et al.*¹ lies.

Ovchinnikov and colleagues sequenced Neanderthal mitochondrial DNA and found that it is closely related, but not identical, to that described previously. Like the first paper², the study of Ovchinnikov *et al.* is convincing in itself. The authors used all the state-of-the-art controls to monitor artefacts and contamination, including having the sequences verified by another laboratory. However, only the combination of the papers allows us to appreciate fully their individual worth. The identification of two Neanderthal DNA sequences, from different specimens found in locations far apart, that are closely related but not identical, rules out the possibility that either sequence is an artefact or the product of contamination. By verifying each other, the two papers provide the most reliable proof so far of the authenticity of ancient DNA sequences.

Can we learn anything from this new Neanderthal DNA sequence about the relationship between modern humans and Neanderthals? The new sequence shares with the Feldhofer one the same surprising feature: it is no more closely related to DNA from modern European populations than to sequences from any other modern human population. This argues against the idea that modern Europeans are at least partly of Neanderthal origin. Although the two sequences were taken from specimens at geographically distant locations, the number of differences between the sequences indicates that these two individuals were from a single gene pool. Furthermore, the variation between the two Neanderthal sequences is similar to that among modern humans.

Details of the Mezmaiskaya sequence also support the suggestion² that there was no contribution of the Neanderthals to the pool of mitochondrial genes in modern human populations. However, this does not exclude the possibility of a contribution of nuclear Neanderthal genes. Approximate quantification of the number of mitochondrial DNA molecules found in the Feldhofer Neanderthal ruled out any hope of recovering nuclear DNA from this specimen², but the apparently excellently preserved Mezmaiskaya specimen might yield values compatible with retrieval of nuclear DNA.

Having achieved DNA sequencing from members of geographically distant Neanderthal populations, it would be interesting to do the same for populations that are far apart on the timescale. A specimen dated closer to the upper time limit of Neanderthal distribution (about 230,000 years ago) would be a tempting choice for DNA retrieval.

The quality of the molecular data retrieved so far from Neanderthal specimens is compelling. If this is how research on ancient DNA is going to proceed, then we are truly on our way to Neanderthal population genetics.

Matthias Höss is at the Swiss Institute for Experimental Cancer Research, Chemin des Boveresses 155, 1066 Epalinges, Switzerland.
e-mail: Matthias.Hoss@isrec.unil.ch

1. Ovchinnikov, I. V. *et al.* *Nature* **404**, 490–493 (2000).
2. Krings, M. *et al.* *Cell* **90**, 19–30 (1997).
3. Golenberg, E. M. *et al.* *Nature* **344**, 656–658 (1990).
4. Woodward, S. R., Weyand, N. J. & Bunnell, M. *Science* **266**, 1229–1232 (1994).
5. DeSalle, R., Gatesy, J., Wheeler, W. & Grimaldi, D. *Science* **257**, 1933–1936 (1992).
6. Cano, R. J., Poinar, H. N., Roubik, D. W. & Poinar, G. O. *J. Med. Sci. Res.* **20**, 619–622 (1992).
7. Cano, R. J., Poinar, H. N., Pieniazek, N. J., Acra, A. & Poinar, G. O. Jr *Nature* **363**, 536–538 (1993).
8. Sidow, A., Wilson, A. C. & Pääbo, S. *Phil. Trans. R. Soc. Lond. B* **333**, 429–432 (1991).
9. Zischler, H. *et al.* *Science* **268**, 1192–1193 (1995).
10. Austin, J. J., Ross, A. J., Smith, A. B., Fortey, R. A. & Thomas, R. H. *Proc. R. Soc. Lond. B* **264**, 467–474 (1997).
11. Walden, K. K. & Robertson, H. M. *Mol. Biol. Evol.* **14**, 1075–1077 (1997).
12. Handt, O. *et al.* *Science* **264**, 1775–1778 (1994).

Daedalus

Direct illumination

The best monochromatic microwave sources are those efficient cavity-resonator valves, the klystron and the magnetron. The wavelength they emit is determined by the size of their cavity. These days, silicon microengineering can create and shape objects and cavities with dimensions around a wavelength of light. So Daedalus is now planning microklystrons and micro-magnetrons to generate bright monochromatic light.

It should be simple enough to form a proper cavity in silicon, and coat it with low-resistance metal to make it a good resonator. The main problem is making an electron beam fine enough to enter the cavity. A conventional cathode would be far too coarse. Fortunately, point sources for electron beams already exist, for field-ion and electron microscopes. With luck they could be formed by the same photofabrication used to make the cavity.

A single photoklystron would be a wonderful point source of light, but very feeble as a lamp. A DREADCO team is designing an array of many hundreds of photoklystrons sculpted on a single silicon wafer. Integrated circuitry on the same wafer will apply control and power voltages to the klystrons of the array. The whole thing will be sealed in an evacuated glass envelope from which the generated light can escape.

Integrated-circuit mass production is so cheap that the photoklystron lamp could well be competitive with its fluorescent and filament rivals. If the klystrons on the wafer were all different, so that each radiated its own frequency, the resulting lamp could be almost perfectly white, ideal for domestic illumination. Even better, it would be a cold white, with no infrared. And as a resonant device, the lamp should be highly efficient, far brighter per watt than its rivals.

If all the photoklystrons in the lamp were identical, it would be perfectly monochromatic, an ideal spectroscopic source. Like a conventional klystron, it could be tuned over a useful range by varying the voltage applied to its electrodes. A series of such klystron lamps could span the visible and infrared bands seamlessly, right out to the millimetre region. Physicists would at last be released from the inflexible, untunable grip of the laser.

David Jones

The Further Inventions of Daedalus (Oxford University Press), 148 past Daedalus columns expanded and illustrated, is now on sale. Special *Nature* offer: m.curtis@nature.com

Finger-length ratios and sexual orientation

Measuring people's finger patterns may reveal some surprising information.

Animal models have indicated that androgenic steroids acting before birth might influence the sexual orientation of adult humans. Here we examine the androgen-sensitive pattern of finger lengths¹, and find evidence that homosexual women are exposed to more prenatal androgen than heterosexual women are; also, men with more than one older brother, who are more likely than first-born males to be homosexual in adulthood², are exposed to more prenatal androgen than eldest sons. Prenatal androgens may therefore influence adult human sexual orientation in both sexes, and a mother's body appears to 'remember' previously carried sons, altering the fetal development of subsequent sons and increasing the likelihood of homosexuality in adulthood.

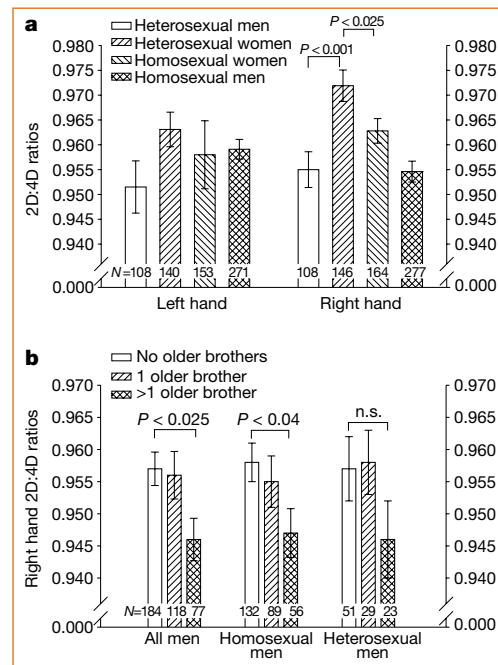
In women, the index finger (2D, second digit) is almost the same length as the fourth digit (4D), although it may be slightly longer or shorter; in men, the index finger is more often shorter than the fourth. The greater 2D:4D ratio in females is established in two-year-olds¹. Because all non-gonadal somatic sex differences in humans appear to be the result of fetal androgens that masculinize males³, the sex difference in the 2D:4D ratio probably reflects the prenatal influence of androgen on males⁴.

In an anonymous survey, 720 adults who were attending public street fairs in the San Francisco area were asked their gender, age, sexual orientation, handedness, and the number and gender of children their mother had carried before them. As expected, men have significantly longer fingers than women ($P < 0.001$), and we confirmed reports that the 2D:4D ratio is greater in women than it is in men.

This sex difference in 2D:4D is greater on the right hand than on the left (Fig. 1a), indicating that the right-hand 2D:4D is more sensitive to fetal androgens than the left-hand ratio. The right-hand 2D:4D ratio of homosexual women was significantly more masculine (that is, smaller) than that of heterosexual women, and did not differ significantly from that of heterosexual men. Thus finger ratios, like otoacoustic emissions⁵, suggest that at least some homosexual women were exposed to greater levels of fetal androgen than heterosexual women.

The 2D:4D ratio of homosexual men was not significantly different from that of heterosexual men for either hand ($P > 0.09$). However, segregating male subjects based on birth order provided support for the role of fetal androgens in male sexual orientation. The more older brothers a boy

Figure 1 Finger-length patterns vary with gender, sexual orientation and birth order. **a**, Among heterosexuals, the mean 2D:4D ratio is larger in women than in men, especially on the right hand. The right-hand 2D:4D ratio of homosexual women is more masculine (that is, smaller) than that of heterosexual women. **b**, Men with more than one older brother are more likely to be homosexual² and have a significantly more masculine right-hand 2D:4D ratio than men without older brothers. Subjects were offered lottery 'scratcher' tickets for their participation. Age and handedness were not significantly different between heterosexual and homosexual subjects; we found no relation between handedness or age and finger measures. Finger lengths were measured blindly from photocopies of subjects' hands. Repeated measurements from 381 subjects were highly correlated ($r = +0.95$ to $+0.99$). Standard errors of the means are depicted. All P values, from Student's t -tests, are two-tailed. n.s., Not significant.



has, the more likely he is to develop a homosexual orientation². Confirming these reports, we also found that only homosexual men had a greater than expected proportion of brothers ($P < 0.01$) among their older siblings (229 brothers: 163 sisters) compared with the general population (106 males: 100 females⁶).

We found that the male 2D:4D ratio, which is unlikely to be influenced by social factors, also varies with the number of older brothers. The ratio was significantly more masculine in men with two or more older brothers than in men with no older brothers (Fig. 1b). There is also a significant correlation ($r = -0.104$; $P < 0.05$) between the number of older brothers and the right-hand 2D:4D ratio in men. If male subjects are divided by sexual orientation, the same pattern of later-born men displaying a more masculine 2D:4D is seen. Having older sisters has no apparent influence on male sexual orientation², or on the 2D:4D ratio in men. No effect of older brothers or sisters on 2D:4D in women was observed, consonant with reports that older siblings exert no effect on female sexual orientation⁷.

Our results suggest that events before birth (or even before conception in the case of older brothers) influence human sexual orientation. The masculinized right-hand 2D:4D ratio in homosexual women may reflect fetal androgen levels that are slightly higher than in heterosexual women. Homosexual men without older brothers have 2D:4D ratios indistinguishable from hetero-

sexual eldest sons, indicating that factors other than fetal androgen (such as genetic influences^{8,9}) also contribute to sexual orientation. Finger measures indicate that men with more elder brothers, including those men who develop a homosexual orientation, might be exposed to greater than normal levels of prenatal androgen.

Although hyper-androgenization of homosexual men might not fit some cultural expectations¹⁰, homosexual men display several hyper-masculine characteristics, including a greater mean number of sexual partners in a lifetime than heterosexual men, who in turn report more sexual partners than do women of either orientation. Furthermore, reports that adult homosexual men have more circulating androgens (ref. 11, but see ref. 12), larger genitalia¹³ and more 'masculine' auditory evoked potentials than heterosexual men¹⁴, are consistent with at least some homosexual men being hyper-androgenized.

Although it is possible that the maternal influence on finger growth of subsequent sons occurs after birth, a prenatal influence seems more likely because of the extensive physiological pairing of mother and fetus. The locus of the maternal 'memory' for previous sons, and the mechanisms by which fetal development of subsequent sons is altered, remain unknown.

Terrance J. Williams, Michelle E. Pepitone, Scott E. Christensen, Bradley M. Cooke, Andrew D. Huberman, Nicholas J. Breedlove, Tessa J. Breedlove, Cynthia L. Jordan,

S. Marc Breedlove

Department of Psychology and Graduate Groups
Neuroscience, Endocrinology, 3210 Tolman Hall,
MC 1650, University of California, Berkeley,
California 94720-1650, USA
e-mail: breedsm@socrates.berkeley.edu

1. Manning, J. T., Scutt, D., Wilson, J. & Lewis-Jones, D. I. *Hum. Reprod.* **13**, 3000–3004 (1998).
2. Blanchard, R. *Annu. Rev. Sex Res.* **8**, 27–67 (1997).
3. Breedlove, S. M., Cooke, B. M. & Jordan, C. L. *Brain Behav. Evol.* **54**, 8–14 (1999).
4. Manning, J. T., Trivers, R. L., Singh, D. & Thornhill, R. *Nature* **399**, 214–215 (1999).

Drosophila cryptochromes

A unique circadian-rhythm photoreceptor

Cryptochrome proteins are critical for circadian rhythms, but their function(s) is uncertain. Here we show that a mutation in a cryptochrome (dCRY) from the fruitfly *Drosophila* blocks an essential photoresponse of circadian rhythms, namely arrhythmicity under constant light conditions. We conclude that dCRY acts as a key photoreceptor for circadian rhythms and that there is probably no other comparable photoreceptor in this species.

Constant light causes the intrinsic circadian period of diurnal animals to shorten and that of nocturnal animals to lengthen (Aschoff's rule). More intense light produces more extreme effects, ultimately resulting in arrhythmicity¹ in most mammals and birds. The circadian period of arthropods generally lengthens in constant light, whether the animal is nocturnal or diurnal¹. *Drosophila melanogaster* is no exception and intense constant illumination leads to arrhythmicity².

The cryptochrome family includes blue-light photoreceptors³. The single known *Drosophila* cryptochrome is thought to be a circadian photoreceptor: flies carrying a mutant allele, *cry^b*, have severely decreased circadian photoresponses⁴, whereas overproduction of dCRY causes increased photosensitivity⁵. In mammals, however, mCRY1 and mCRY2 are more likely to be involved in the central clock mechanism^{6–8}.

This raises the possibility that dCRY effects on photosensitivity reflect a role downstream of circadian photoreception, somewhere along the circadian light-input pathway or within the *Drosophila* central clock itself. This fits with the fact that *cry^b* flies are still able to reset their circadian rhythm (entrain) to new light–dark cycles². We find, however, that *cry^b* flies remain behaviourally rhythmic in intense constant light, in contrast to wild-type flies and many other species which are arrhythmic under such conditions (Fig. 1a,b)^{1,2}.

The arrhythmicity of *cry^b* flies must be a

5. McFadden, D. & Pasanen, E. *Proc. Natl Acad. Sci. USA* **95**, 2709–2713 (1998).
6. James, W. H. *Hum. Biol.* **59**, 721–752 (1987).
7. Bogaert, A. F. *Behav. Neurosci.* **111**, 1395–1397 (1997).
8. Bailey, J. M. & Pillard, R. C. *Arch. Gen. Psychiatry* **48**, 1089–1096 (1991).
9. Hamer, D. D., Hu, S., Magnuson, V. L., Hu, N. & Patautucci, A. M. L. *Science* **261**, 321–327 (1993).
10. Gorman, M. R. *Persp. Biol. Med.* **38**, 61–81 (1994).
11. Brodie, H. K. H. et al. *Am. J. Psychiatry* **131**, 82–83 (1974).
12. Mayer-Bahlburg, H. F. L. *Progr. Brain Res.* **61**, 375–398 (1984).
13. Bogaert, A. F. & Hershberger, S. *Arch. Sexual Behav.* **28**, 213–221 (1999).
14. McFadden, D. & Champlin, C. A. *J. Ass. Res. Otolaryngol.* (in the press).

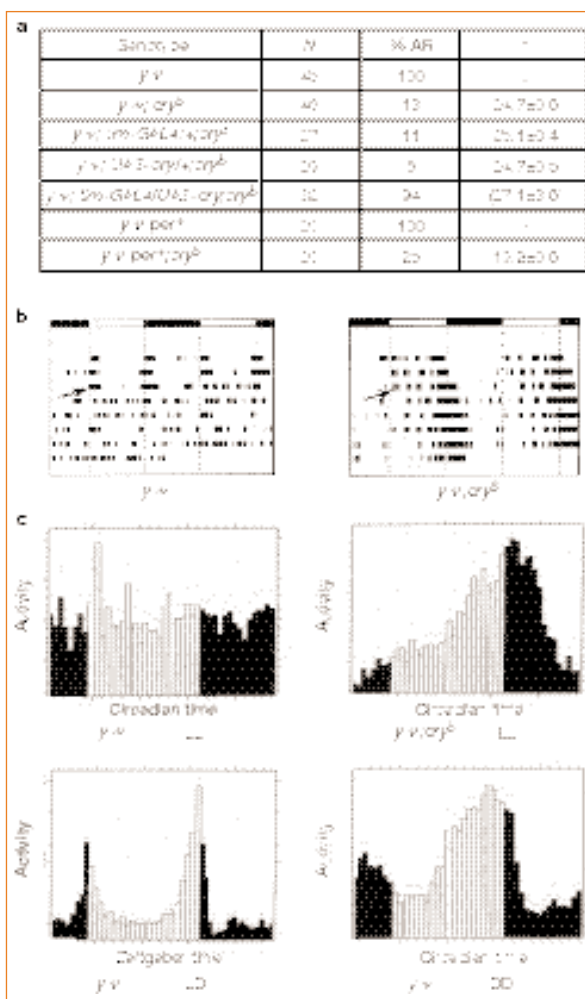
property of the *cry* gene, because the normal phenotype can be rescued by expressing wild-type dCRY in rhythm-generating cells of *cry^b* flies. In intense constant light, the *cry^b* mutant's behaviour is strikingly similar to that of wild-type flies in constant darkness (Fig. 1c). An identical 24.7-hour period is also recorded under constant-darkness conditions, indicating that this slightly longer period is a characteristic of the background genotype. Thus, there is not even a detectable lengthening of period in the *cry^b* mutant strain under constant light condi-

tions². The free-running, circadian nature of *cry^b* flies' constant-light behaviour is further demonstrated by the 19.2-h period observed for flies with *cry^b* in combination with a short allele of the *period* gene (*per^s*; Fig. 1a).

Our results show that the *cry^b* mutation impairs the circadian photoreception pathway so profoundly that the fly cannot 'see' constant light. This mutant also responds very poorly to short light pulses⁴; by these criteria, this circadian photoreceptor must be unique in *Drosophila*. How then can *cry^b* flies entrain to different 24-h light–dark cycles? The missense *cry^b* mutation might generate a protein with weak activity that would be sufficient for light–dark entrainment but not for a normal arrhythmic behavioural response to constant illumination. However, our previous results suggest a different explanation: entrainment of *cry^b* flies is through a second, completely separate light-input pathway⁴. Visual photoreception^{4,9} may even directly influence locomotor activity, which then affects circadian rhythms only indirectly through a non-photoc phase-resetting pathway.

Our results indicate that dCRY is an important circadian photoreceptor and

Figure 1 *cry^b* circadian rhythms free-run under intense constant light. **a**, *cry^b* behavioural rhythmicity and rescue of wild-type arrhythmicity. Flies were entrained under a normal 12-h light:12-h dark regime for 3 d. At the end of the fourth light period, the light was left on at saturating light intensities (2,000 lux) and activity monitored over the next 6 d. The last 5 d were used to analyse locomotor activity rhythms. *tim-GAL4* and *UAS-cry* are two transgenes⁵ used to express wild-type CRY in circadian-rhythm-generating cells of *cry^b* flies. *N*, number of flies analysed; % AR, percentage of arrhythmic flies (cut off, 'power' <10 or 'width' <2; ref. 10); τ , period length of the circadian behavioural rhythms in hours. Value in parentheses is based on only two weakly rhythmic flies. **b**, Representative *y w* and *y w;cry^b* actograms; data are double-plotted. After the light is permanently left on (arrow), *cry^b* fly behaviour starts to free-run, whereas wild-type flies become arrhythmic. The first day of entrainment is not shown. White boxes, days or subjective days; black boxes, nights or subjective nights. **c**, Average activity plots of 14 *y w* and 16 *y w;cry^b* flies under constant light conditions (LL), and 16 *y w* flies under light–dark (LD) or constant-darkness conditions (DD). White bars, days or subjective days (zeitgeber or circadian time, 0–12); black bars, nights or subjective nights (zeitgeber or circadian time, 12–24). Each bar represents the 5-day average activity of a pool of flies during a specific 30-min period of the day (for example, the first white bar of an LD plot is the average activity of a pool of flies between zeitgeber time 0 and 0.5 during the 5 d of measurement). Dots, standard deviation.



probably the only dedicated one in *Drosophila*. Although we cannot yet exclude additional central-clock functions for dCRY, the abrogation of constant-light effects in *cry^b* mutant flies indicates that this cryptochrome makes a unique contribution to *Drosophila* circadian photoreception.

Patrick Emery*, Ralf Stanewsky†, Jeffrey C. Hall*, Michael Rosbash*†

*Department of Biology, NSF Center for Biological Timing, †Howard Hughes Medical Institute, Brandeis University, Waltham, Massachusetts 02454, USA

‡Zoologisches Institut, Universität Regensburg, D93040 Regensburg, Germany
e-mail: rosbash@brandeis.edu

1. Aschoff, J. Z. *Tierpsychol.* **49**, 225–249 (1979).
2. Konopka, R. J., Pittendrigh, C. & Orr, D. J. *Neurogenet.* **6**, 1–10 (1989).
3. Cashmore, A. R., Jarillo, J. A., Wu, Y. J. & Liu, D. *Science* **284**, 760–765 (1999).
4. Stanewsky, R. et al. *Cell* **95**, 681–692 (1998).
5. Emery, P., So, W. V., Kaneko, M., Hall, J. C. & Rosbash, M. *Cell* **95**, 669–679 (1998).
6. Kume, K. et al. *Cell* **98**, 193–205 (1999).
7. van der Horst, G. T. et al. *Nature* **398**, 627–630 (1999).
8. Vitaterna, M. H. et al. *Proc. Natl Acad. Sci. USA* **96**, 12114–12119 (1999).
9. Ohata, K., Nishiyama, H. & Tsukahara, Y. in *Biological Clocks: Mechanisms and Applications* (ed. Touitou, Y.) 167–170 (Elsevier, Amsterdam, 1999).
10. Ewer, J., Frisch, B., Hamblen-Coyle, M. J., Rosbash, M. & Hall, J. C. *J. Neurosci.* **12**, 3321–3349 (1992).

Structural colour

Colour mixing in wing scales of a butterfly

Green coloration in the animal kingdom, as seen in birds' feathers and reptile integument, is often an additive mixture of structurally effected blue and pigmentary yellow¹. Here we investigate the origin of the bright green coloration of the wing scales of the Indonesian male *Papilio palinurus* butterfly, the microstructure of which generates an extraordinary combination of both yellow and blue iridescence. The dual colour arises from a modulation imposed on the multilayer, producing the blue component as a result of a previously undiscovered retro-reflection process.

Scanning electron micrographs of scales taken from the wings' coloured regions show that their surfaces comprise a regular two-dimensional array of concavities, of about 4–6 µm in diameter and 0.5–3 µm at the greatest depth. Transmission electron micrographs of these scales in cross-section reveal the multilayering that causes the iridescence as well as a modulation that leads to retro-reflection (Fig. 1a).

The variation in colour across each concave surface modulation is evident from optical microscopy. In reflection, for normally incident light, the flat regions between and

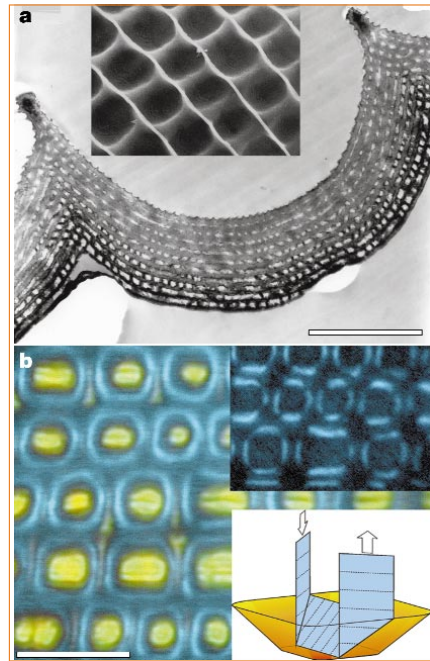


Figure 1 Modulated multilayering leads to dual colour in *P. palinurus*. **a**, Transmission electron microscope image showing a cross-section through one concavity on a *P. palinurus* iridescent scale. Inset; a scanning electron microscope image of the surface of an iridescent scale. Scale bar, 1 µm (inset, 7 µm). **b**, Real-colour image showing the dual-colour nature of the reflectivity from the surface of the *P. palinurus* iridescent scale, taken using unpolarized light in an optical microscope. Top inset, image of the same region taken with crossed polarizers. Bottom inset, illustration of the mechanism by which polarization is converted through double reflection from orthogonal sides of a concavity. Scale bar, 12 µm (inset, 6 µm).

in each concavity appear yellow, and the inclined sides of each concavity appear blue (Fig. 1b). It is the juxtaposition of these yellow and blue regions that synthesizes the green coloration perceived by the human eye, as they are too small to be resolved individually. Such spatial-averaging colour-stimulus synthesis² has been reported in beetles³, and it also forms the basis of colour-television pictures and pointillistic painting.

The blue component cannot be back-reflected from a single multilayer system inclined at 45° to the incident direction. The effect is in fact caused by a pair of orthogonal multilayer surfaces that lie on opposite sides of each concavity. Light incident along the scale perpendicular, reflected from one surface inclined at 45°, is directed across the concavity to the opposite, orthogonal surface, where it is returned back along the incident direction. These pairs of inclined surfaces with almost identical multilayering have matched spectral reflectivity characteristics; this causes intense blue reflectivity through this double reflection.

Support for this retro-reflection mechanism comes from evidence of polarization conversion in the reflection from these scales. When we cross an input linear polarizer with an exit analyser while viewing the sample

under normally incident light, all yellow reflected light is extinguished, but a substantial amount of blue reflected light remains, indicating that the blue reflected light has undergone polarization conversion. This effect occurs after double reflection from a pair of orthogonal surfaces when the wavevector of the incident light is at 45° to the reflecting surface and the polarization vector is at 45° to the plane of incidence.

Under diffuse white light, humans see the wing's green coloration in a limited solid angle about the wing normal. Outside this perspective, the wing colour changes predictably, becoming bluer as observation approaches grazing incidence⁴. The retro-reflection from pairs of opposite sides of each concavity is then less effective because their angle-dependent spectral reflectivity characteristics become mismatched. However, increasingly non-normal incidence observation is facilitated through large-angle reflections from the bottom and single sides of each concavity.

The purpose of this mechanism of colour generation is unclear. Structural colours can provide higher visibility⁵ than pigmentary colours and can, given the appropriate microstructure⁶, create colour-dependent polarization and angle effects. Conspecific and predator photoreceptor sensitivity must also be considered. Species whose spectral-vision sensitivity spans the two reflected structural colours may perceive a third by colour-stimulus synthesis. Polarization sensitivity associated with such photoreceptors⁷ would provide further detail from wing reflectivity about species type and even wing orientation.

The mechanism by which *P. palinurus* produces its bi-colour and polarization effects is optically rare (although similar but less pronounced polarization effects and colour-stimulus synthesis of green have been identified in the other *Papilio* butterflies, *P. crino*, *P. buddha* and *P. blumei*). Through simple modulation of an otherwise uniform multilayer system, it synthesizes a very different colour stimulus in certain visual systems. The structure shows strong local polarization conversion of only one of the colours, through the mechanism we term orthogonal-surface retro-reflection.

P. Vukusic*, J. R. Sambles*, C. R. Lawrence†

*Thin Film Photonics, School of Physics, University of Exeter, Exeter EX4 4QL, UK

†Mechanical Sciences Sector, DERA, Farnborough, GU14 0LX, UK

e-mail: P.Vukusic@ex.ac.uk

1. Fox, D. L. *Animal Biochromes and Structural Colours* (Univ. California Press, Berkeley, 1976).
2. Burnham, R. W., Hanes, R. M. & Bartleson, C. J. *Color* (Wiley, New York, 1963).
3. Schultz, T. D. & Bernard, G. D. *Nature* **337**, 72–73 (1989).
4. Huxley, J. *J. Entomol. A* **50**, 9–22 (1975).
5. Vukusic, P., Sambles, J. R., Lawrence, C. R. & Wootton, R. J. *Proc. R. Soc. Lond. B* **266**, 1402–1411 (1999).
6. Ghiradella, H. *Microsc. Anat. Invert.* **A 11**, 257–287 (1998).
7. Kelber, A. *Nature* **402**, 251 (1999).

Resolving the extragalactic hard X-ray background

R. F. Mushotzky*, L. L. Cowie†, A. J. Barger† & K. A. Arnaud*‡

* NASA Goddard Space Flight Center, Code 662, Greenbelt, Maryland 20771, USA

† Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA

‡ Astronomy Department, University of Maryland, College Park, Maryland 20742, USA

The origin of the hard (2–10 keV) X-ray background has been a mystery for over 35 years. Most of the soft X-ray background has been resolved into individual sources (mainly quasars), but these sources do not have the spectral energy distribution required to match the spectrum of the X-ray background as a whole. Here we report the results of a deep survey, using the Chandra satellite, in which the detected hard X-ray sources account for at least 75 per cent of the hard X-ray background. The mean X-ray spectral energy distribution of these sources is in good agreement with that of the background. Moreover, most of those hard X-ray sources are associated unambiguously with either the nuclei of otherwise normal bright galaxies, or with optically faint sources. The latter could be active nuclei in dust-enshrouded galaxies or a population of quasars at extremely high redshift.

For some time after the discovery of the cosmic X-ray background (XRB¹), there was considerable controversy over whether the background arose from a superposition of discrete sources or from thermal bremsstrahlung emission from a hot intergalactic gas. We now know that the bulk of the XRB cannot originate in a uniform, hot intergalactic medium because a strong Compton distortion on the cosmic microwave background spectrum was not observed by the FIRAS instrument on the COBE satellite^{2,3}.

At soft X-ray energies (0.5–2 keV), the XRB has been extensively studied with the ROSAT satellite. The deepest ROSAT source counts reach ~1,000 per square degree at a limiting flux of 10^{-16} erg cm⁻² s⁻¹, and at this level 70–80% of the XRB is resolved into discrete sources⁴. Most of the optical identifications of a complete sample of 50 ROSAT sources, at a limiting flux of 5×10^{-15} erg cm⁻² s⁻¹, are unobscured active galactic nuclei (AGN)⁵. But because the objects detected in the soft band do not have the spectrum of the XRB, a new population of absorbed or flat spectrum objects are needed to make up the background at higher energies. Detailed models developed to resolve this ‘spectral paradox’ assumed that most of the flux in the XRB is produced by active galaxies that are obscured by dust. When deep imaging sky surveys with the ASCA^{6–9} and BeppoSAX¹⁰ satellites became possible in the hard (>2 keV) X-ray band, ~30% of the hard XRB was resolved, but only indirect identifications of the optical counterparts could be made.

The Chandra satellite¹¹, with its great sensitivity over a wide energy range, excellent image quality, superb positional accuracy, and reasonable field-of-view, can directly image the sources that make up the hard XRB. We have therefore carried out a deep imaging survey of the Hawaii Deep Survey Field SSA13 with the ACIS-S instrument on Chandra to resolve the hard XRB and to identify the nature of the sources that produce it. We chose to centre on the SSA13 field¹², which has existing multiwavelength observations^{13–15}, to maximize the immediate identification of optical/near-infrared counterparts and redshifts for the X-ray source detections. We find that above a flux threshold of 2.5×10^{-15} erg cm⁻² s⁻¹ (2–10 keV), we can account for at least 75% of the sky flux, with the main uncertainty being the sky flux itself. Our deep optical observations show a rich assortment of hard X-ray sources which could not have been discovered by previous satellites.

Chandra X-ray survey of SSA13

The SSA13 observation was performed on 3–4 December 1999 for an elapsed time of 100.9 ks. The optical axis of the telescope at right

ascension (2000) 13 h 12 min 21.40 s, declination (2000) 42° 41′ 20.96″ was positioned on the back-illuminated CCD (S3) of ACIS-S because this detector has a much better soft X-ray sensitivity than the front-illuminated chips. Furthermore, since the back-illuminated detectors did not suffer the radiation damage which affected the front-illuminated chips in orbit, they are well characterized by extensive ground-based calibrations.

The overall sensitivity of the instrument spans a wide energy range, from 0.2 to 10 keV. Two energy-dependent images of the S3 chip were generated in the hard (2–10 keV) and soft (0.5–2 keV) bands, as was a 2–10 keV image of the front-illuminated S2 chip that covered a neighbouring region. We extracted sources independently for the hard- and soft-band images. Sources brighter than 3.2×10^{-15} erg cm⁻² s⁻¹ (2–10 keV) or 3×10^{-16} erg cm⁻² s⁻¹ (0.5–2 keV; S3 chip only) which lie within 6 arcmin of the optical axis are given in Table 1, ordered by right ascension; the table contains 22 sources selected in the hard band, and a further 15 sources selected solely in the soft band. Details of the extraction and calibration of the X-ray data and of the optical photometry may be found in Table 1 legend.

Number counts and the X-ray background

The cumulative counts per square degree, $N(>S)$, are the sum of the inverse areas of all sources brighter than flux S . Sources at the faintest fluxes can be detected only at smaller off-axis angles where the point-spread function and vignetting corrections are smaller; thus, the area diminishes with flux. In Fig. 1a and b we present our cumulative counts per square degree (filled squares) in the soft and hard bands, respectively, with 1σ uncertainties from the Poisson error in the number of detected sources (jagged solid lines). To the limiting flux levels of 2.3×10^{-16} erg cm⁻² s⁻¹ (0.5–2 keV) and 2.5×10^{-15} erg cm⁻² s⁻¹ (2–10 keV), simulations show that the counts are nearly complete and that Eddington bias is unimportant; thus, the raw counts accurately represent the true counts.

Our soft-band counts are in excellent agreement with the deep ROSAT counts in the Lockman hole from Hasinger *et al.*⁴ in the region of overlap. At fainter fluxes our new counts fall at the lower limit of their fluctuation analysis, which suggests a continuing flattening.

An area-weighted maximum likelihood fit¹⁶ of a single power law to the 0.5–2 keV counts over the flux range (2.3–70) $\times 10^{-16}$ erg cm⁻² s⁻¹ is given by the relation

$$N(>S) = 185(S/7 \times 10^{-15})^{-0.7 \pm 0.2} \quad (1)$$

where the errors on the power-law index are 68% confidence. Likewise, a power-law fit to the 2–10 keV counts over the flux range $(2.5\text{--}20) \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$ is given by

$$N(>S) = 170(S/2 \times 10^{-14})^{-1.05 \pm 0.35} \quad (2)$$

where the counts intercept the ASCA extrapolation at the upper end of the flux range. Though the range in indices is consistent with the power-law index of 1.5 seen at brighter fluxes, the counts are significantly lower than an extrapolation of the ASCA counts.

The source contributions to the XRB can be obtained by summing the individual fluxes divided by area or, more indirectly, by integrating SdN using the power-law fits. We list the directly summed source contributions to the XRB in the two bands in Table 2, along with previous determinations by ROSAT and ASCA. With the additional 10% contribution from our data to the soft band, a maximum flux of $1.1 \times 10^{-12} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ deg}^{-2}$ remains to be accounted for. In the hard band, the combination of the present results with the ASCA measurements at higher fluxes means

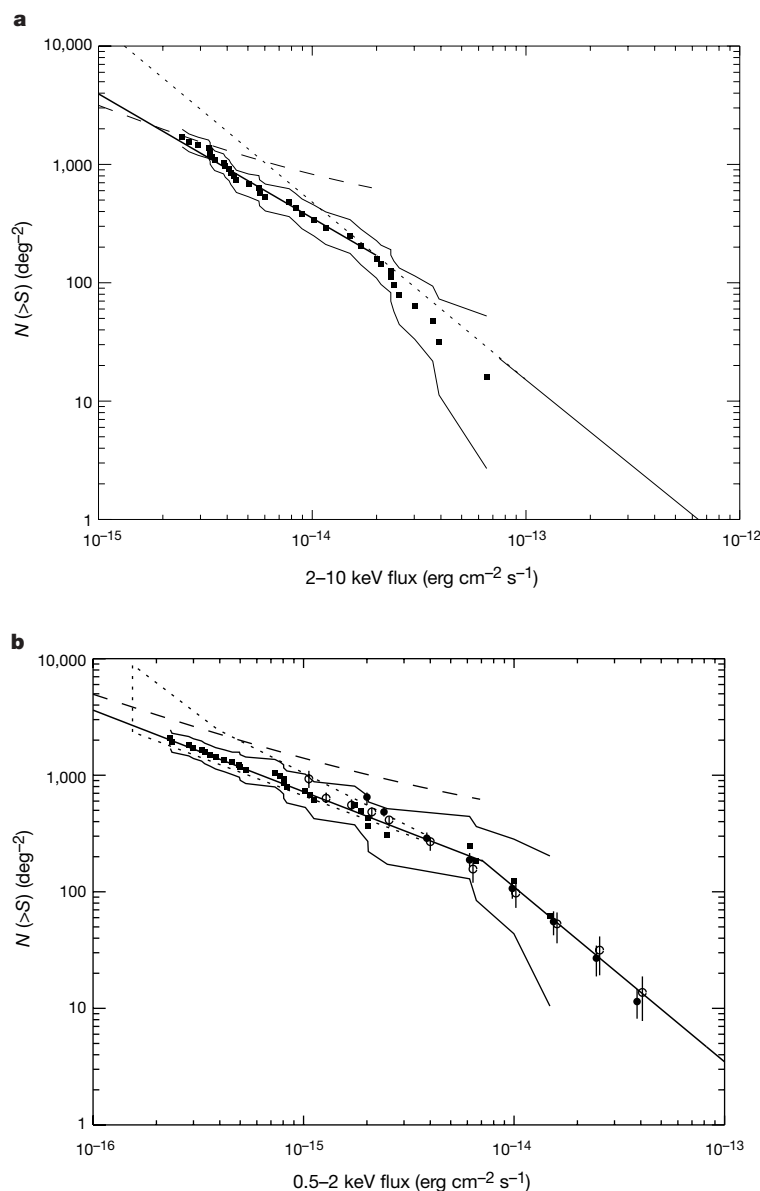


Figure 1 Integral number counts per square degree of X-ray sources in the SSA13 field versus flux for the hard (a) and the soft (b) energy bands. These data are shown by the squares. The counts in the soft band are based on 30 sources in the 10^{-7} probability sample covering an area of 59 square arcminutes on the S3 chip. The hard counts are based on the 10^{-7} probability sample of the S2 chip in the 2–10 keV band and a 10^{-7} probability sample of the S3 chip chosen in the 2–6 keV band (to minimize background) and corrected to 2–10 keV fluxes. The total area is 84 square arcminutes. At fluxes greater than $2 \times 10^{-14} \text{ erg cm}^{-2} \text{ s}^{-1}$, sources were drawn from the S2, S3, I2 and I3 chips, which increased the area to 227 square arcminutes. The combined hard sample contains 35 sources. In a, the solid line at bright fluxes is the $N(>S) \propto S^{-3/2}$ representation of the ASCA counts from Ueda *et al.*⁸ (sensitivity limit $7 \times 10^{-14} \text{ erg cm}^{-2} \text{ s}^{-1}$); these data lie on the extrapolation from previous results by

HEAO1 A2⁴³ with a euclidean slope of -1.5 . The dotted line shows the extrapolation of this line to fainter fluxes. The solid line at fainter fluxes is the -1.05 power-law fit to the present data below $2 \times 10^{-14} \text{ erg cm}^{-2} \text{ s}^{-1}$. The dashed line shows the normalization at a given flux at which integral counts with the observed shape would exceed a 2–10 keV sky flux²⁰ of $1.9 \times 10^{-11} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ deg}^{-2}$. In b, the open and filled circles show the counts determined from the ROSAT PSPC camera (sensitivity limit $2 \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$) and HRI camera (sensitivity limit $10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$) data of the Lockman hole from Hasinger *et al.*⁴ The dotted line shows the fluctuation limits from Hasinger *et al.*²⁵ The solid lines are the -0.7 index power-law fit to the data below $7 \times 10^{-14} \text{ erg cm}^{-2} \text{ s}^{-1}$ and the -1.5 index power-law fit above that flux. The dashed line shows the normalization at a given flux at which integral counts with the observed shape would exceed the 0.5–2 keV XRB²⁰.

that at least 75% of the background (using the highest published normalization) is resolved to the currently observed flux limits.

Optical properties of the X-ray sources

We have compared our X-ray images with existing^{17,18} deep HK' , I , B and U' images obtained with the Keck 10-m and University of Hawaii 2.2-m telescopes. Because of the excellent ~ 1 -arcsec X-ray positional accuracy, we can, in most cases, securely identify the optical counterparts to the X-ray sources. In Fig. 2 we show thumbnail I -band images of all of the X-ray sources in Table 1. Only one source, CXO J131159.3+423928 (significant in both the hard and soft X-ray images), is significantly extended in the X-ray images; it is probably a high-redshift cluster. The optical image (no. 35 in Fig. 2) is centred on a faint ($I = 23$ mag) galaxy which lies at the centre of a region of enhanced galaxy density. In addition to the probable cluster, the hard sample contains two quasars, eight bright galaxies, and ten optically faint ($I > 23$ mag) objects, while the soft sample contains five quasars, three bright galaxies, and seventeen optically faint objects. Morphologically, the X-ray selected bright galaxy population consists of a mixture of early spirals and elliptical galaxies. Three of the bright galaxies show possible signs of inter-

action with nearby bright neighbours while the remainder are clearly isolated.

Our Keck spectra for the quasars show broad $Mg II$ or $C IV$ and Lyman- α lines. In some of the bright galaxy spectra clear AGN signatures are present (for example, a broad absorption line galaxy with P-Cygni profiles at redshift $z = 1.320$). However, although subtle AGN signatures may be present in their optical spectra, most of the bright galaxies would not have been identified as AGN in an optical survey. We illustrate this in Fig. 3, with spectra for three of the bright galaxies.

The X-ray spectrum

The photon intensity of the XRB, $P(E)$, where E is the photon energy in keV and $P(E)$ has units of photons $cm^{-2} s^{-1} keV^{-1} sr^{-1}$, can be approximated by a power law, $P(E) = AE^{-\Gamma}$. The A-2 experiment¹⁹ on the HEAO1 satellite found that the XRB spectrum from 3 to 15 keV was well described by a photon index $\Gamma \approx 1.4$, and this result has been confirmed and extended to lower energies by recent analyses of ASCA^{20–23} and BeppoSAX²⁴ data.

The photon indices of the individual sources given in Table 1 were computed from the ratios of the counts in the 0.5–2 keV band to

Table 1 X-ray sources

No.	RA (2000) (h min s)	Dec. (2000) (° ' ")	$f(2-10 \text{ keV})$ (10^{-16} cgs)	$f(0.5-2 \text{ keV})$ (10^{-17} cgs)	Γ	HK' (mag)	I (mag)	B (mag)	U' (mag)	z
0*	13 12 43.38	42 44 36.73	82.5 ± 21.0	ND	ND	ND	20.33	ND	ND	0.463
1†	13 12 40.24	42 39 35.55	43.0 ± 14.1	79.0 ± 16.7	0.93	ND	24.43	25.19	ND	ND
2*	13 12 39.62	42 45 48.77	89.9 ± 25.2	ND	ND	ND	>26.7	ND	ND	ND
3*	13 12 39.50	42 42 48.83	39.2 ± 11.4	ND	ND	ND	17.89	ND	ND	0.111
4‡	13 12 37.94	42 40 5.53	32.1 ± 11.8	32.8 ± 10.8	0.45	ND	24.10	25.76	ND	ND
5*	13 12 37.16	42 43 21.08	50.0 ± 12.5	ND	ND	ND	>26.7	ND	ND	ND
6†	13 12 36.86	42 38 44.55	46.2 ± 14.2	39.2 ± 12.1	0.28	ND	>25.4	>26.7	ND	ND
7†	13 12 36.58	42 40 2.80	384 ± 34.0	1480 ± 67.1	1.54	ND	>25.4	26.70	ND	ND
8§	13 12 36.00	42 40 44.11	41.6 ± 12.3	23.1 ± 9.22	-0.06	ND	23.94	25.40	>25.5	ND
9†	13 12 35.68	42 41 50.67	232 ± 26.0	408 ± 35.3	0.90	ND	20.91	22.95	23.40	1.320
10*	13 12 34.48	42 43 9.27	78.1 ± 14.9	ND	ND	16.38	19.26	23.22	24.19	0.241
11‡	13 12 32.36	42 39 49.39	15.7 ± 8.86	50.0 ± 12.9	1.38	20.16	24.37	25.82	>25.5	ND
12§	13 12 31.34	42 39 2.19	48.4 ± 13.1	19.8 ± 8.60	-0.40	18.02	20.54	23.37	23.76	0.585
13‡	13 12 30.83	42 39 42.73	4.71 ± 6.69	38.2 ± 11.3	2.12	19.75	23.94	25.80	25.36	ND
14‡	13 12 29.26	42 37 32.33	14.5 ± 10.6	112 ± 20.1	2.09	16.04	18.39	20.91	21.35	0.234
15*	13 12 28.25	42 44 54.52	56.7 ± 13.7	ND	ND	21.64	24.41	25.33	>25.5	ND
16†	13 12 26.00	42 37 35.86	49.9 ± 15.2	170 ± 24.6	1.43	19.73	22.67	24.53	25.14	1.426
17‡	13 12 25.29	42 41 19.53	13.7 ± 8.14	49.6 ± 12.8	1.45	21.68	>25.9	26.93	>25.5	ND
18¶	13 12 22.48	42 44 49.97	42.1 ± 11.7	ND	ND	20.73	24.98	26.26	>25.5	ND
19†	13 12 22.32	42 38 13.89	116 ± 19.8	622 ± 44.6	1.80	17.80	19.84	21.53	22.23	2.565#
20‡	13 12 21.63	42 35 49.97	45.4 ± 28.4	203 ± 33.9	1.65	ND	25.06	>26.7	ND	ND
21‡	13 12 21.50	42 44 5.41	17.3 ± 8.92	78.7 ± 16.1	1.60	19.23	22.06	23.36	23.98	1.305#
22†	13 12 20.11	42 42 22.42	49.0 ± 12.3	196 ± 24.6	1.53	>22.5	24.14	25.75	>25.5	ND
23‡	13 12 19.19	42 38 8.36	28.3 ± 11.5	36.1 ± 11.6	0.62	>22.5	24.56	26.72	>25.5	ND
24†	13 12 15.32	42 39 0.22	190 ± 23.3	986 ± 54.4	1.75	16.18	17.92	18.66	19.00	2.565#
25‡	13 12 11.72	42 44 12.59	19.5 ± 10.0	175 ± 24.3	2.16	18.65	20.70	22.22	22.27	0.950#
26†	13 12 10.02	42 41 29.94	151 ± 20.3	246 ± 27.8	0.76	16.74	19.47	22.50	24.15	0.212
27‡	13 12 9.93	42 36 15.30	15.9 ± 25.8	77.3 ± 22.2	1.72	19.66	23.50	27.32	>25.5	ND
28‡	13 12 8.38	42 41 43.08	2.37 ± 6.18	53.1 ± 13.4	2.88	20.39	23.24	25.64	>25.5	ND
29†	13 12 6.55	42 41 41.31	138 ± 20.2	95.8 ± 17.7	0.05	17.80	20.50	23.66	24.82	0.696
30§	13 12 6.55	42 41 25.16	38.4 ± 11.8	8.17 ± 6.45	-1.72	15.22	17.47	20.52	22.31	0.180
31‡	13 12 5.18	42 41 23.45	7.29 ± 7.77	33.8 ± 11.3	1.62	21.16	>25.9	28.02	>25.5	ND
32§	13 12 4.18	42 41 13.59	48.2 ± 13.7	25.9 ± 10.2	-0.17	22.39	>25.9	>27.6	>25.5	ND
33‡	13 12 1.23	42 42 7.78	5.28 ± 8.64	83.3 ± 17.9	2.60	21.08	23.65	26.67	>25.5	3.405#
34‡	13 11 59.66	42 41 52.89	34.3 ± 13.9	42.1 ± 13.5	0.53	ND	>25.9	26.21	>25.5	ND
35†	13 11 59.36	42 39 28.17	273 ± 36.5	612 ± 51.2	1.01	ND	23.04	>27.6	>25.5	ND
36‡	13 11 59.19	42 38 34.12	55.5 ± 22.6	107 ± 23.3	0.93	ND	>25.9	27.29	>25.5	ND

Shown are data for X-ray sources in the SSA13 field selected in the hard (2–10 keV; S2 and S3 chips) or soft (0.5–2 keV; S3 chip) bands. The X-ray images were prepared using xselect and associated tools at GSFC. ACIS grades 0, 2, 3, 4 and 6 were used, and columns at the boundaries of the readout nodes were rejected. Counts lying within a $5''$ diameter aperture were measured, together with the background in a $5''$ – $7.5''$ radius annulus, at $2''$ intervals along the field. The distribution of detected counts is poissonian. A cut of 17 counts in the hard S3 image and 10 counts in the other two images represents a $<10^{-7}$ probability threshold against background fluctuations, and ensures a $<20\%$ probability of a single spurious source detection in the entire sample. The source counts were corrected for the enclosed energy fraction within the aperture. For the S3 chip, the flux calibrations were made using an array of effective areas versus energy at 12 positions and an assumed power-law spectrum having counts-weighted mean photon indices $\Gamma = 1.2$ (2–10 keV) and $\Gamma = 1.4$ (0.5–2 keV). The galactic $N(H) = 1.4 \times 10^{20} \text{ cm}^{-2}$ is too low to affect the flux conversions. For the S2 chip a single conversion factor of $2.6 \times 10^{-11} \text{ erg cm}^{-2} \text{ count}^{-1}$ was used. Using the on-axis flux calibrations of 2.5×10^{-11} (2–10 keV) and 2.9×10^{-12} (0.5–2 keV) to convert the S3 counts per second to flux, we determine limiting minimum fluxes of $3.2 \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$ (2–10 keV) and $3.0 \times 10^{-16} \text{ erg cm}^{-2} \text{ s}^{-1}$ (0.5–2 keV). The table is restricted to sources with off-axis angles $<6''$, where $>50\%$ of the energy is enclosed within a $\sim 2.5''$ radius, and to sources where the noise, computed from the variance of the background and signal, is less than one-third of the signal. The $15''$ borders of each chip have incomplete exposure times due to the spacecraft dither so objects detected in these borders were not included in our counts analysis. The near-infrared and optical magnitudes are computed in $1.5''$ radii intervals. I is Kron-Cousins, B is Johnson, HK' is a broad filter centred at $1.9 \mu\text{m}$, and U' is a $300\text{-}\text{\AA}$ filter centred at $3,400\text{\AA}$. Lower limits are 1σ . ND, not determined.

* Detected in the S2 chip.

† Significant detection in both samples.

‡ S3 source detected only in soft-band sample.

§ S3 source detected only in hard-band sample.

|| In the S3 $15''$ excluded border.

¶ In the S2 $15''$ excluded border.

Quasar spectrum.

those in the 2–10 keV band, assuming that each source could be described by a single power law. There is an extremely wide range of hardness in both samples, ranging from negative indices to $\Gamma = 1.8$ in the hard-selected sample, and from $\Gamma = 0.1$ to values above 2 in the soft sample. The composite (counts-weighted) photon index is 1.22 ± 0.03 in the hard sample and 1.42 ± 0.04 in the soft sample. The progressive hardening of the soft sample as we move to fainter fluxes is a continuation of a trend seen in the ROSAT samples^{25,26}. The combined spectrum of all the soft X-ray sources of Table 1 is well fitted by a single power law over the 0.3–10 keV range with an index of 1.42 ± 0.07 and an extinction corresponding to the galactic $N(H) = 1.4 \times 10^{20} \text{ cm}^{-2}$. If we assume that 75% of the 2–10 keV background has an index of 1.22, and that the remaining 25% of the background comes from sources that have fluxes greater than $1 \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$ (2–10 keV) and an average photon index⁷ of 1.63, then the index of the combined sample is 1.38, which agrees extremely well with the spectrum of the hard X-ray background^{19–24}.

Inspection of Table 1 suggests that the hardest sources tend to correspond to the bright galaxies, with the optically faint objects having intermediate hardness, and the quasars being the softest of the sources observed. We illustrate this more clearly in Fig. 4, where we have plotted *I*-band magnitude versus photon index. Of the objects with $I \approx 22$ mag, more than half are galaxies, and the majority of these have $\Gamma < 1$. The five known quasars all lie in the $\Gamma > 1.7$ range, consistent with that of most brighter AGN. The faint sources spread over a wide range of indices that overlap both of the other populations. We can quantify this by generating the counts-weighted averages for each population separately. For the hard-selected sample, we find that the bright galaxies (nos 9, 12, 26, 29) have an average photon index of 0.59 ± 0.06 , the faint objects (nos 1, 6, 7, 8, 9 and 22) have 1.33 ± 0.06 , and the two quasars have

Table 2 Source contributions to the X-ray background

Energy range (keV)	Flux range ($\text{erg cm}^{-2} \text{ s}^{-1}$)	Source contributions ($\text{erg cm}^{-2} \text{ s}^{-1} \text{ deg}^{-2}$)	Percentage of XRB*	Reference
1–2	$>10^{-15}$	3.0×10^{-12}	68–81	Hasinger <i>et al.</i> ⁴
1–2	$(2.3\text{--}10) \times 10^{-16}$	$(3.8 \pm 1.0) \times 10^{-13}$	6–13	This work
2–10	$>10^{-13}$	4.5×10^{-12}	20–28	Ueda <i>et al.</i> ⁷
2–10	$(2.5\text{--}100) \times 10^{-15}$	$(1.30 \pm 0.3) \times 10^{-11}$	56–81	This work

The statistical errors on our observed sky brightnesses dominate the systematic errors, which are expected to be less than 10%. To be consistent with Hasinger *et al.*⁴, we converted our soft-band sky brightness of $(6.0 \pm 1.5) \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ deg}^{-2}$ to the 1–2 keV range using the measured mean photon index. This result was then compared with the 1–2 keV background (where galactic contamination is less than at lower energies) of $(3.7\text{--}4.4) \times 10^{-12} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ deg}^{-2}$ from Gendreau *et al.*²¹ using a fit to ASCA data, and Chen *et al.*²⁰, using a fit to joint ASCA/ROSAT data. In the hard band the summed counts are compared with the 2–10 keV background of $(1.6\text{--}2.3) \times 10^{-11} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ deg}^{-2}$ from Marshall *et al.*¹⁹, using a fit to HEAO1 A2 data, and Vecchi *et al.*²⁴, using a fit to BeppoSAX data. XRB, X-ray background.

*The range given is a combination of the uncertainty in the source contribution (from the third column) and the variation in the published sky flux.

1.76 ± 0.07 . For the soft-selected sample, the 16 objects with $I > 23$ mag have a composite index of 1.35 ± 0.06 , which is almost identical to that of the optically faint objects in the hard sample, and the quasars have a composite index of 1.80 ± 0.12 .

The source of the background

Our data conclusively show that AGN are the main contributors to the hard X-ray background. Many of our sources agree with the predictions of XRB synthesis models^{27–33}; these models were constructed within the framework of AGN unification schemes to account for the spectral intensity of the hard XRB, and to explain the X-ray source counts in the hard and soft energy bands. In the unified scheme, the orientation of a molecular torus surrounding the nucleus determines the classification of the source. The models

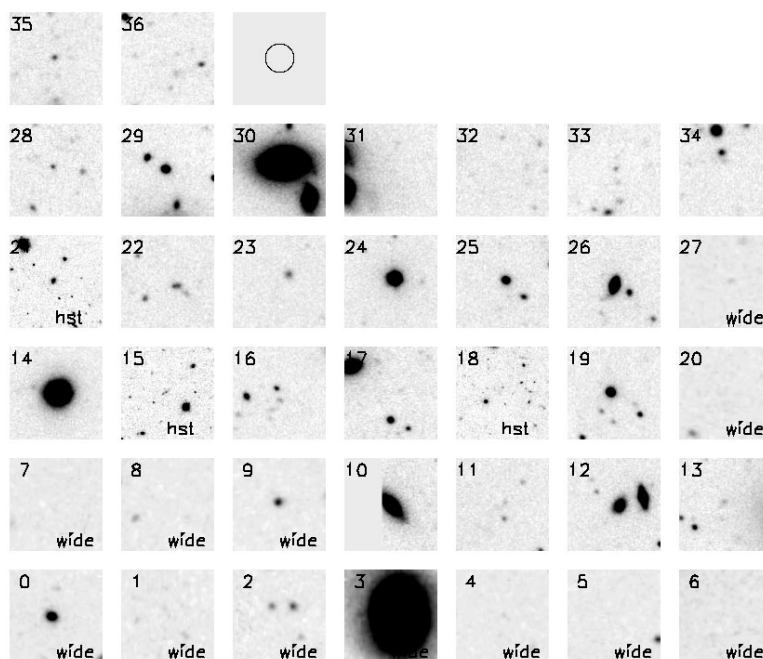


Figure 2 *I*-band images $9'' \times 9''$ of the X-ray sources of Table 1. Most of the images are from ultradeep data obtained with the Low Resolution Imaging Spectrograph (LRIS) on the Keck 10-m telescope, but those marked wide are shallower wide-field data from the University of Hawaii 2.2-m telescope, and those marked hst are from deep (approximately 16,000 s of exposure) F814W Hubble Space Telescope data. The astrometry of the optical images is tied to deep 20-cm Very Large Array images currently being analysed by E. Richards *et al.* (manuscript in preparation). The absolute offset from the nominal Chandra astrometry (2.2'' W, 0.2'' N) was obtained from the quasar CXO J131215.3+423901. A small adjustment to the pixel size (0.4908'' versus 0.492'') was

also made to optimize the agreement between the X-ray and optical sources, but no adjustment of the roll angle. In this system the r.m.s. dispersion between the 8 objects in the soft sample with $I < 23$ mag optical counterparts in the deep LRIS data is 0.36''. Intercomparison of the independent positions determined from the hard and soft images suggests that the error may rise to as much as 1.7'' in the faintest X-ray sources. The identification numbers are as in Table 1, and the sources are ordered from the lower left by right ascension. The rightmost panel in the top row shows a circle of 1.5'' radius, typical of the maximum positional uncertainty.

invoke, along with a population of unobscured AGN (whose nuclear emission we see directly), a substantial population of intrinsically obscured AGN whose hydrogen column densities (of $N_{\text{H}} \approx 10^{21}\text{--}10^{25} \text{ cm}^{-2}$) around the nucleus block our line-of-sight.

The AGN that make up the hard XRB come in two main types: roughly 40% are luminous early-type galaxies (both ellipticals and early spirals) in the redshift range from $z = 0$ to just beyond $z = 1$, and roughly 50% have faint or, in some cases, undetectable, optical counterparts. Most of these objects would not have been found even in sensitive optical surveys for AGN.

The bright galaxy population is extremely X-ray hard, with an average photon index of $\Gamma = 0.59$. The X-ray sources are point-like and centred on the galaxy nuclei, which suggests that they are produced by accretion onto the central black holes that are known to be present in such systems. The hardness of the X-ray spectra indicates that these X-ray sources are highly obscured. Such sources were described by Moran *et al.*³⁶ on the basis of data from the Einstein satellite. After hard X-ray components were discovered by Allen *et al.*³⁸ in ASCA spectra of six nearby giant elliptical galaxies, a

model was constructed which was able to account for a large fraction of the XRB with objects of this type. The model predicted that a significant fraction of the hard number counts at fluxes $< 10^{-14} \text{ erg cm}^{-2} \text{ s}^{-1}$ could arise from sources at low redshift, as is indeed now observed to be the case. The absolute K magnitudes of these sources lie between ~ -24 and ~ -26 ($H_0 = 65 \text{ km s}^{-1} \text{ Mpc}^{-1}$ and $q_0 = 0.5$), or from just below to several times the L_* luminosity, where L_* is the characteristic galaxy luminosity, and their rest-frame 2–10 keV luminosities range from 5×10^{41} to $3 \times 10^{43} \text{ erg s}^{-1}$. These sources are at too low redshifts to be likely submillimetre candidates; however, they should be far-infrared sources, which SIRTf and other forthcoming airborne and space missions should be able to detect.

The optically faint sources have an average photon index of $\Gamma = 1.3$. These sources could either be a smooth continuation to $z > 1$ of the bright early-type galaxies with obscured luminous X-ray nuclei, more distant obscured AGN, or something more exotic, such as extremely high redshift ($z \gg 5$) quasars. For this final possibility, the objects would be invisible in the B band because of scattering by the foreground intergalactic neutral hydrogen. In the soft sample, 11 of the optically faint sources have $B > 26 \text{ mag}$ and could lie in this category. This places an upper limit on the density of objects on the sky of this type of source of 0.26 per square arcminute to the $3 \times 10^{-16} \text{ erg cm}^{-2} \text{ s}^{-1}$ limit of the 0.5–2 keV sample, which is slightly lower than the predictions of the model of Haiman and Loeb³⁹ for X-ray selected high-redshift quasars. The handful of objects which are detected in the near-infrared but absent in B band are the most promising candidates for this type of object and, in some cases (for example, object 27 of Table 1) may be bright enough for follow-up with near-infrared spectroscopy to test the hypothesis.

Comparison with submillimetre^{40–42}, far-infrared and radio samples should allow us to determine what fraction of objects in these surveys are X-ray-emitting AGN. As near-infrared spectra and photometric redshift estimates of the optically faint sources are established, we should be able to refine the obscured AGN models and determine whether any of the faint sources are indeed very-high-redshift quasars. With the X-ray, optical and submillimetre samples all now approaching the full resolution of their respective

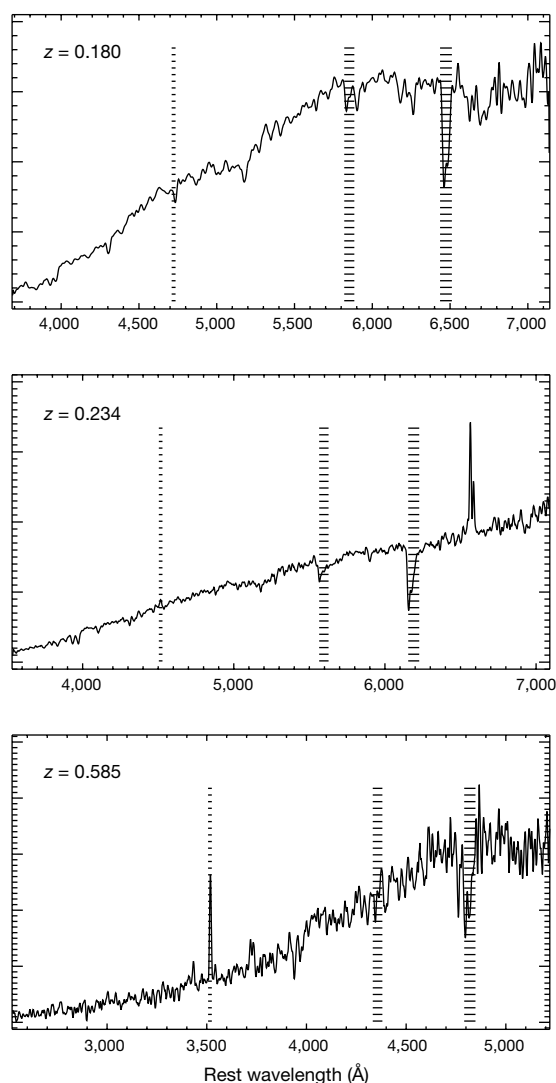


Figure 3 Keck LRIS spectra of three of the X-ray-detected bright galaxies whose redshifts have been determined. These are source no. 30 at $z = 0.180$ and source 12 at $z = 0.585$ from the hard-selected sample and source 14 from the soft sample. These objects do not show strong emission features, except for H α , N II and S II in the $z = 0.234$ spectrum and [O II] 3,727 Å in the $z = 0.585$ spectrum. The resolution of the spectra is 14 km s^{-1} and the shaded regions show the positions of the strong 5,577-Å night sky line and the atmospheric bands.

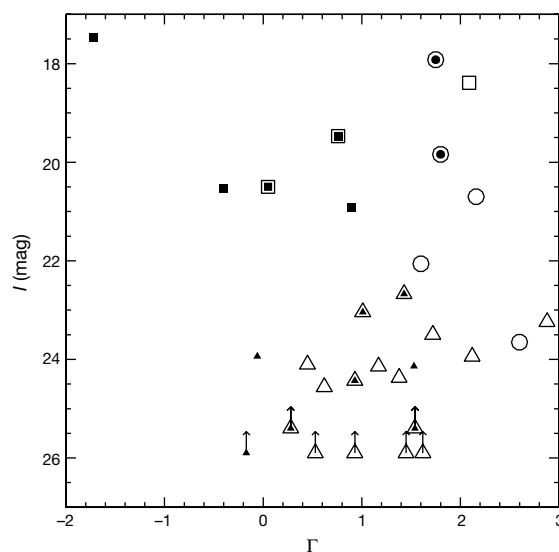


Figure 4 I-band magnitudes versus X-ray photon indices. Filled symbols represent the hard (2–10 keV) selected sample, and open symbols represent the soft (0.5–2 keV) selected sample. For objects with redshift identifications, quasars are represented by circles and galaxies by squares. Sources without redshift identifications are represented by triangles. In cases where the objects were significantly detected in both samples, the hard-selected magnitudes and indices were used; these objects are indicated by filled symbols surrounded by larger open symbols.

backgrounds, we are close to achieving a complete cosmic census of the population of galaxies and AGN. □

Received 30 December 1999; accepted 15 February 2000.

- Giacconi, R., Gursky, H., Paolini, F. & Rossi, B. Evidence for X-rays from sources outside the solar system. *Phys. Rev. Lett.* **9**, 439–443 (1962).
- Mather, J. C. *et al.* Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. *Astrophys. J.* **420**, 439–444 (1994).
- Wright, E. L. *et al.* Interpretation of the COBE FIRAS CMBR spectrum. *Astrophys. J.* **420**, 450–456 (1994).
- Hasinger, G. *et al.* The ROSAT deep survey. I. X-ray sources in the Lockman field. *Astron. Astrophys.* **329**, 482–494 (1998).
- Schmidt, M. *et al.* The ROSAT deep survey. II. Optical identification, photometry and spectra of X-ray sources in the Lockman field. *Astron. Astrophys.* **329**, 495–503 (1998).
- Ueda, Y. *et al.* A population of faint galaxies that contribute to the cosmic X-ray background. *Nature* **391**, 866–868 (1998).
- Ueda, Y. *et al.* Log N–log S relations and spectral properties of sources from the ASCA large sky survey: their implications for the origin of the cosmic X-ray background (CXB). *Astrophys. J.* **518**, 656–671 (1999).
- Ueda, Y., Takahashi, T., Ishisaki, Y., Ohashi, T. & Makishima, K. The ASCA medium sensitivity survey (the GIS catalog project): source counts and evidence for emerging population of hard sources. *Astrophys. J.* **524**, L11–L14 (1999).
- Cagnoni, L., Della Ceca, R. & Maccacaro, T. A medium survey of the hard X-ray sky with the ASCA gas imaging spectrometer: The (2–10 keV) number counts relationship. *Astrophys. J.* **493**, 54–61 (1998).
- Fiore, F. *et al.* The contribution of faint AGN to the hard X-ray background. *Mon. Not. R. Astron. Soc.* **306**, L55–L60 (1999).
- Weisskopf, M. C. The Chandra X-ray Observatory (CXO): an overview. In *Proc. NATO ASI 'The Neutron Star–Black Hole Connection'* (in the press); also as preprint astro-ph/9912097 at (<http://xxx.lanl.gov>) (1999).
- Lilly, S. J., Cowie, L. L. & Gardner, J. P. A deep imaging and spectroscopic survey of faint galaxies. *Astrophys. J.* **369**, 79–105 (1991).
- Songaila, A., Cowie, L. L., Hu, E. M. & Gardner, J. P. The Hawaii K band galaxy survey III. Spectroscopy of $K < 20$ galaxies. *Astrophys. J. Suppl. Ser.* **94**, 461–515 (1994).
- Windhorst, R. A. *et al.* Identification of faint radio sources with optically luminous interacting disk galaxies. *Nature* **375**, 471–474 (1995).
- Barger, A. J., Cowie, L. L. & Sanders, D. B. Resolving the submillimeter background. *Astrophys. J.* **518**, L5–L8 (1999).
- Murdoch, H. S., Crawford, D. F. & Jauncey, D. L. Maximum likelihood estimation of the number-flux density distribution of radio sources in the presence of noise and confusion. *Astrophys. J.* **183**, 1–13 (1973).
- Cowie, L. L., Songaila, A., Hu, E. M. & Cohen, J. G. New insight on galaxy formation and evolution from Keck spectroscopy of the Hawaii deep fields. *Astron. J.* **112**, 839–864 (1996).
- Wilson, G., Cowie, L. L., Barger, A. J. & Burke, D. J. The evolution of the universal ultraviolet luminosity density with redshift. *Astrophys. J.* (submitted).
- Marshall, F. *et al.* The diffuse X-ray background spectrum from 3 to 50 keV. *Astron. J.* **235**, 4–10 (1980).
- Chen, L.-W., Fabian, A. C. & Gendreau, K. C. ASCA and ROSAT observations of the QSF3 field: the X-ray background in the 0.1–7 keV band. *Mon. Not. R. Astron. Soc.* **285**, 449–471 (1997).
- Gendreau, K. C. *et al.* ASCA observations of the spectrum of the X-ray background. *Publ. Astron. Soc. Jpn* **47**, L5–L9 (1995).
- Miyaji, T. *et al.* The cosmic X-ray background spectrum observed with ROSAT and ASCA. *Astron. Astrophys.* **334**, L13–L16 (1998).
- Ishisaki, Y. *et al.* *Astrophys. J.* (submitted).
- Vecchi, A., Molendi, S., Guainazzi, M., Fiore, F. & Parmar, A. N. The BeppoSAX 1–8 keV cosmic background spectrum. *Astron. Astrophys.* **349**, L73–L76 (1999).
- Hasinger, G. *et al.* A deep X-ray survey in the Lockman hole and the soft X-ray log N–log S. *Astron. Astrophys.* **275**, 1–15 (1993).
- Almaini, O. *et al.* A deep ROSAT survey. XII. The X-ray spectra of faint ROSAT sources. *Mon. Not. R. Astron. Soc.* **282**, 295–303 (1996).
- Setti, G. & Woltjer, L. Active galactic nuclei and the spectrum of the X-ray background. *Astron. Astrophys.* **224**, L21–L23 (1989).
- Madau, P., Ghisellini, G. & Fabian, A. C. The unified Seyfert scheme and the origin of the cosmic X-ray background. *Mon. Not. R. Astron. Soc.* **270**, L17–L21 (1994).
- Matt, G. & Fabian, A. C. Spectral constraints on Seyfert 2 galaxies as major contributors to the hard (3–100 keV) X-ray background. *Mon. Not. R. Astron. Soc.* **267**, 187–192 (1994).
- Comastri, A., Setti, G., Zamorani, G. & Hasinger, G. The contribution of AGNs to the X-ray background. *Astron. Astrophys.* **296**, 1–12 (1995).
- Zdziarski, A. A., Johnson, W. N., Done, C., Smith, D. & McNaron-Brown, K. The average X-ray/gamma-ray spectra of Seyfert galaxies from GINGA and OSSE and the origin of the cosmic X-ray background. *Astrophys. J.* **438**, L63–L66 (1995).
- Smith, D. A. & Done, C. Unified theories of active galactic nuclei: a hard X-ray sample of Seyfert 2 galaxies. *Mon. Not. R. Astron. Soc.* **280**, 355–377 (1996).
- Gilli, R., Rinaliti, G. & Salvati, M. Beyond the standard model for the cosmic X-ray background. *Astron. Astrophys.* **347**, 424–433 (1999).
- Schmidt, M., Giacconi, R., Hasinger, G., Trüper, J. & Zamorani, G. The X-ray luminosity function of active galactic nuclei. In *Highlights in X-ray Astronomy* (Max-Planck Institut für Extraterrestrische Physik report, in the press); also as preprint astro-ph/9908295 at (<http://xxx.lanl.gov>) (1999).
- Miyaji, T., Hasinger, G. & Schmidt, M. Soft X-ray AGN luminosity function from ROSAT surveys I. *Astron. Astrophys.* (in the press); also as preprint astro-ph/9910410 at (<http://xxx.lanl.gov>) (1999).
- Moran, E. C., Helfand, D. J., Becker, R. H. & White, R. L. The Einstein two-sigma catalog: silver needles in the X-ray haystack. *Astrophys. J.* **461**, 127–145 (1996).
- Allen, S. W., Di Matteo, T. & Fabian, A. C. Hard X-ray emission from elliptical galaxies. *Mon. Not. R. Astron. Soc.* (in the press); also as preprint astro-ph/9905052 at (<http://xxx.lanl.gov>) (1999).
- Di Matteo, T. & Allen, S. W. Hard X-ray emission from elliptical galaxies and its contribution to the X-ray background. *Astrophys. J.* **527**, L21–L24 (1999).
- Haiman, Z. & Loeb, A. X-ray emission from the first quasars. *Astrophys. J.* **521**, L9–L13 (1999).
- Hughes, D. H. *et al.* High-redshift star formation in the Hubble deep field revealed by a submillimetre-wavelength survey. *Nature* **394**, 241–247 (1998).
- Barger, A. J. *et al.* Submillimetre-wavelength detection of dusty star-forming galaxies at high redshift. *Nature* **394**, 248–251 (1998).
- Gunn, K. F. & Shanks, T. Implications of an obscured AGN model for the X-ray background at submillimetre and far-infrared wavelengths. *Mon. Not. R. Astron. Soc.* (submitted); also as preprint astro-ph/9909089 at (<http://xxx.lanl.gov>) (1999).
- Piccinotti, G. *et al.* A complete X-ray sample of the high latitude ($|b| > 20^\circ$) sky from HEAO 1 A-2: log N–log S and luminosity functions. *Astrophys. J.* **253**, 485–503 (1982).

Acknowledgements

We thank J. Halpern and G. Hasinger for comments which improved the first draft of this Article; we also thank the CXC, L. Van Speybroeck, M. Weisskopf and the MSFC team, M. Bautz, G. Garmire, and the ACIS team for building and operating the observatory. We acknowledge the use of HEASARC software. A.J.B. was supported by the Hubble and Chandra fellowship programs.

Correspondence and requests for materials should be addressed to L.L.C. (e-mail: cowie@ifa.hawaii.edu).

Peptide exosite inhibitors of factor VIIa as anticoagulants

Mark S. Dennis, Charles Eigenbrot, Nicholas J. Skelton, Mark H. Ultsch, Lydia Santell, Mary A. Dwyer*, Mark P. O'Connell & Robert A. Lazarus

Department of Protein Engineering, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA

Potent anticoagulants have been derived by targeting the tissue factor–factor VIIa complex with naive peptide libraries displayed on M13 phage. The peptides specifically block the activation of factor X with a median inhibitory concentration of 1 nM and selectively inhibit tissue-factor-dependent clotting. The peptides do not bind to the active site of factor VIIa; rather, they work by binding to an exosite on the factor VIIa protease domain, and non-competitively inhibit activation of factor X and amidolytic activity. One such peptide (E-76) has a well defined structure in solution determined by NMR spectroscopy that is similar to the X-ray crystal structure when complexed with factor VIIa. These structural and functional studies indicate an allosteric ‘switch’ mechanism of inhibition involving an activation loop of factor VIIa and represent a new framework for developing inhibitors of serine proteases.

Initiation of the extrinsic pathway of the coagulation cascade occurs upon vascular damage, exposing circulating factor VII/VIIa (FVII/FVIIa) to its membrane-bound co-factor, tissue factor (TF)^{1,2}. The TF–FVIIa complex cleaves its zymogen substrates (FX, FIX and FVII) to their corresponding active serine protease forms, which leads to the formation of thrombin and ultimately a fibrin clot³. Despite many years of research using various strategies^{4,5}, only two anticoagulants are in widespread clinical use—coumarins⁶ and heparins^{7,8}. Coumarins impair the function of the vitamin-K-dependent proteins including both procoagulants (thrombin, FXa, FIXa and FVIIa) and anticoagulants (activated protein C and protein S) whereas heparins enhance inhibition of thrombin and FXa by antithrombin III (ATIII). The non-selective mode of inhibition of both of these anticoagulants probably accounts for their therapeutic limitations in maintaining the balance between thrombosis and haemostasis⁴.

Because it initiates the coagulation process, the TF–FVIIa complex represents a good target for developing therapeutic anticoagulants. Human TF–FVIIa is inhibited by tissue factor pathway inhibitor (TFPI)—a protein containing three Kunitz domains that inhibits TF–FVIIa in an FXa-dependent manner⁹—and the serpin ATIII, although the latter is controversial^{10,11}. These proteins are similar to many other naturally occurring serine protease inhibitors that exert their influence by interfering with the catalytic triad at the active site by either reversible or irreversible mechanisms^{12,13}. Small molecule active-site inhibitors¹⁴ have been developed as anticoagulants based upon further understanding of coagulation proteases and their inhibitors^{15,16}.

The elaborate nature of the TF–FVIIa complex suggests additional approaches that may impair its function^{17–19}. As well as the 254-residue serine protease domain in the carboxy-terminal heavy chain, the 152-residue amino-terminal light chain of FVIIa contains two epidermal growth factor (EGF)-like domains and a γ -carboxy glutamic-acid-rich (Gla) domain, which is involved in binding to phospholipid membranes¹⁹. The 2.0 Å crystal structure of the TF–FVIIa complex elucidates the extended conformation of FVIIa and describes the intricate association between the two fibronectin type III domains of TF and the light chain of FVIIa²⁰. Thus, compounds that block the association of TF and FVIIa, prevent macromolecular substrate interaction or interfere with binding to membrane

provide alternative approaches to covalent and reversible active-site inhibitors; examples include antibodies against TF or FVIIa, TFAA (a TF mutant with reduced co-factor function for FX activation) and TF-derived peptides²¹. Here we present a new class of peptide inhibitors of FVIIa, discovered using phage display of naive peptide libraries²², which bind to a distinct and functionally relevant exosite on the TF–FVIIa complex. Structural and functional studies show that these inhibitors possess a unique mechanism of action and represent a new paradigm for developing inhibitors of serine proteases.

Selection and characterization of peptide inhibitors

Polyvalent phage libraries of naive 20-residue peptides constrained by a single disulphide bond²³ were sorted against the immobilized TF–FVIIa complex. After four rounds of selection and amplification, a single clone from the E library emerged (designated E-56). Phage displaying E-56 bound specifically to immobilized FVIIa or TF–FVIIa, but not to TF or bovine serum albumin (BSA). E-56 phage also bound to TF–FVIIa when the active site was alkylated with biotinylated EGR chloromethylketone or blocked by TF71-C, a Kunitz domain inhibitor²⁴. Thus, E-56 binds to FVIIa, but at a site

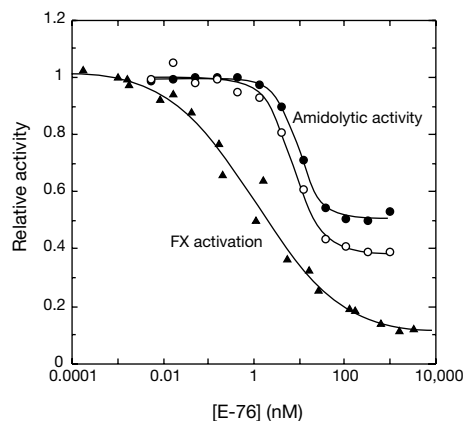


Figure 1 Inhibition of TF–FVIIa-catalysed activation of FX and chromogenic activity by E-76. The IC_{50} values for inhibition of FX activation (triangles) and amidolytic activity (filled circles) were 1.1 nM and 9.7 nM, respectively, as determined by nonlinear regression analysis of the data fit to a four-parameter equation⁴⁴; the IC_{50} value for inhibition of amidolytic activity in the absence of TF_{1–219} (open circles) was 7.4 nM.

* Present address: Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, USA.

distinct from the active site—an 'exosite'. To reduce avidity effects and select for higher affinity sequences, a partially randomized monovalent phage library, which introduced a 50% mutation rate at each position, was made and sorted against TF–FVIIa. This process identified residues that were 100% conserved and therefore likely to be critical for binding. The remaining positions were then completely randomized in additional phage libraries to obtain a 'fully matured' consensus sequence; peptide E-76 (Ac-ALCDDPRVDR-WYQFVEG-NH₂) was chosen for further characterization.

Surface plasmon resonance studies established that E-76 associates with FVIIa or TF–FVIIa with the same dissociation constant (K_d) of around 8.5 nM. No evidence for binding was detected for TF or a number of other serine proteases including thrombin, FXa, FIXa, FXIa, FXIIa, plasma kallikrein, activated protein C, tissue plasminogen activator (t-PA) and plasmin. Furthermore, binding to FVIIa was dependent on the presence of calcium; FVIIa contains Ca²⁺-binding sites in the Gla and EGF1 domains as well as a single site in the protease domain²⁰.

To assess functional activity, we tested E-76 in several TF–FVIIa activity assays. Most importantly, E-76 is a potent inhibitor of FX activation with a median inhibitory concentration (IC₅₀) of 1 nM (Fig. 1). Furthermore, although E-76 does not bind at the active site, it does inhibit the amidolytic activity towards a small chromogenic substrate; this inhibition is independent of TF, consistent with the binding data above (Fig. 1). Kinetic analyses of Lineweaver–Burk plots indicate that E-76 is a non-competitive inhibitor of FX, reducing the maximum reaction velocity (V_{max}) but not altering the Michaelis constant (K_m), and a non-competitive, mixed-type inhibitor of amidolytic activity, affecting both V_{max} and K_m (Fig. 2)²⁵. Inhibition of the latter activity was incomplete, with activity

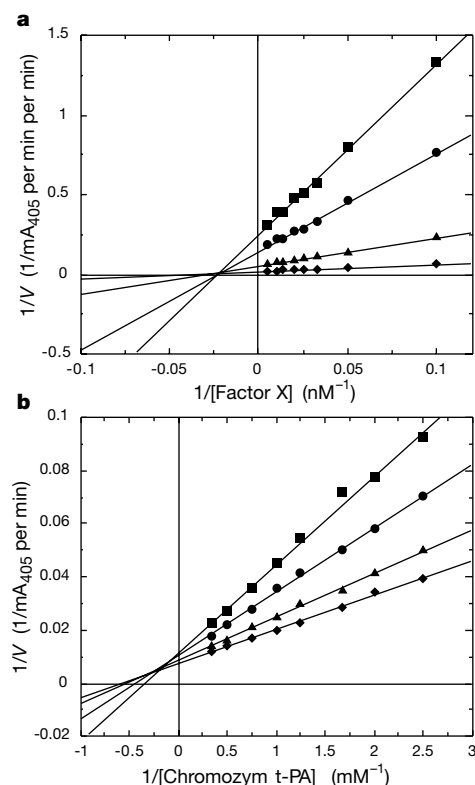


Figure 2 Lineweaver–Burk plots for E-76 inhibition of FX activation and amidolytic activity. **a**, FX activation showed noncompetitive inhibition. Squares, 400 nM E-76; circles, 80 nM; triangles, 16 nM; and diamonds, uninhibited. **b**, Noncompetitive mixed type inhibition was found for Chromozym t-PA turnover. Squares, 200 nM E-76; circles, 22 nM; triangles, 7.4 nM; and diamonds, uninhibited. Lines represent data fit to a linear equation (Kaleidagraph).

reduced by about 50% at saturating concentrations (Fig. 1). TF significantly enhances amidolytic activity²⁶, presumably by allosterically inducing a conformational change in FVIIa^{17–19}. However, binding and subsequent inhibition of FVIIa by E-76 is independent of any TF-induced conformational state (Fig. 1). This implies that this exosite on FVIIa is both topologically distinct and allosterically unlinked from TF.

Consistent with the *in vitro* inhibition data, E-76 is a very potent anticoagulant. Inhibition of the TF-dependent extrinsic clotting pathway, as measured by the concentration-dependent prolongation of the prothrombin time, is shown in Fig. 3. Significantly, there is no evidence for prolonging the clotting time in the TF-independent intrinsic pathway, as determined by the activated partial thromboplastin time (Fig. 3). This indicates that E-76 does not inhibit any of the serine proteases involved in the intrinsic pathway, which include thrombin, FXa, FIXa, FXIa, FXIIa and plasma kallikrein, in agreement with the binding data described above. Obtaining specificity for inhibitors that bind at the active site is inherently difficult owing to the myriad of serine proteases and their similarity at this site. In this regard inhibitors that bind at an exosite may have a distinct advantage. The specificity of E-76 for FVIIa is highly desirable as the therapeutic limitations of other known anticoagulants probably result from their non-selective modes of inhibition⁴.

Peptide structure–activity relationships

Analysis of ¹H NMR data indicates that residues within the disulphide-bonded loop of peptide E-76 have a well defined structure in solution, consisting of a distorted type I reverse turn (Asp 5P–Val 8P) and an irregular turn of helix (Asp 9P–Cys 13P) (E-76 peptide residues are designated with a P following the residue number). Well defined structure in peptides of this size is not common. The termini of E-76 are less well defined, although a C-terminal extension of the helix is partially populated, and Leu 2P does make contact with residues in the helix. A prominent feature of this fold is the cluster of hydrophobic side chains (Leu 2P, Trp 11P, Tyr 12P and Phe 15P) on one face of the peptide (Fig. 4a).

FVIIa binding affinities and inhibition of FX activation were measured for an extensive series of peptide analogues of E-76; the results indicated that the surface containing this hydrophobic cluster contacts FVIIa (Fig. 4b). The NMR spectra for the analogues

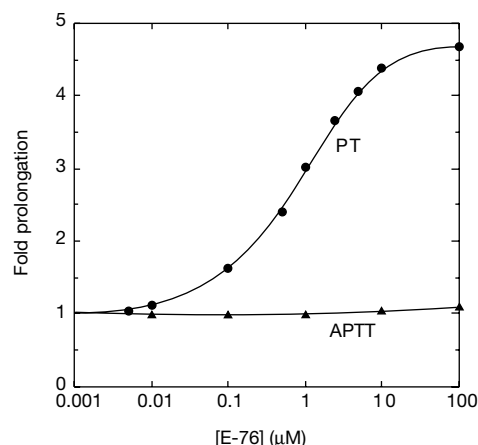


Figure 3 Fold prolongation of prothrombin time (PT) and activated partial thromboplastin time (APTT) of E-76 in human plasma. PT and APTT clotting time assays were performed in pooled normal human plasma using an ACL 300 Automated Coagulation Analyzer. Peptides were diluted 10-fold into citrated plasma and incubated for 30 min before initiation of clotting with Innovin (PT) or Actin FS (APTT); clotting times in the absence of E-76 were 10.2 and 28.1 s, respectively. Lines represent smoothed curves fit to the data (Kaleidagraph).

exhibiting decreased inhibition of FX activation were examined for evidence of structural changes based on inter-residue nuclear Overhauser effects (NOEs) and $^3J_{\text{HN-H}\alpha}$. By these criteria, Val 8PAla and Trp 11PAla were unaltered in the tertiary structure and Leu 2PAla and Phe 15PAla had only local changes (loss of long-range, N-terminal contacts or fraying of the helix C-terminus, respectively). As the structure of Tyr 12PAla was severely perturbed, we analysed the properly folded Tyr 12P Phe analogue. Compared to E-76 this was around 300-fold less potent in inhibiting FX activation, indicating that Tyr 12P is an important FVIIa contact.

Location of the peptide-binding exosite on FVIIa

The E-76 binding site on FVIIa was located using a three-step process of increasing biochemical resolution. Initially, we localized the binding site on FVIIa to the 254-residue serine protease domain, as E-peptide phage bound to FVIIa or a FVII_{protease}/FIX chimaera, but not to a FVII/FIX_{protease} chimaera. As E-76 phage bind to human but not rabbit FVIIa (data not shown) and these protease domains share 70% identity, the binding site was refined by measuring E-76 phage binding to 14 'rabbitized' mutants of human FVIIa encompassing 70 of the 76 sequence differences. Reduced binding was found for only 2 of the 14 mutants and, as these residues were proximal, indicated a likely peptide-binding region. We made individual alanine mutants in this region of human FVIIa to define the E-peptide-binding exosite further. None of the mutants

Table 1 Refinement statistics for Gla-domainless FVIIa with E-76

Refinement resolution (Å)	50.0–3.0
Number of reflections	16915
Number of reflections (R_{free})	654
R	22.5%
R_{free}	29.5%
R (all)	22.8%
r.m.s.d. bond distances (Å)	0.012
r.m.s.d. bond angles (°)	2.0
r.m.s.d. dihedral angles (°)	26.3
r.m.s.d. improper dihedral angles (°)	0.82
Number of atoms	5919
Number of atoms occupancy zero	381
Number of residues	756
Number of waters	4
Number of calcium ions	4
Number of cacodylate ions	2
r.m.s.d. main chain B-factors (Å ²)	3.2
r.m.s.d. side chain B-factors (Å ²)	5.0
Ramachandran plot	
residues in most favourable region (%)	75.8
nonglycine residues in disallowed regions (%)	0.0

except Leu 144 Ala (see below) altered the specific activity for FX activation, implying that they were all properly folded and functional. However, subsequent differences between mutant and wild-type IC₅₀ values for E-76 inhibition of FX activation allowed us to identify the specific binding site. These differences are shown on the structure of FVIIa (Fig. 5) and indicate that the E-peptide exosite is in a trough that is distinct from, but close to, the active site. One wall of this trough contains the protease Ca²⁺-binding site, consistent with the Ca²⁺-dependence of E-peptide binding and inhibition.

Crystal structure of FVIIa complexed with E-76

To assess the structural and functional information obtained from FVIIa binding site mutations and peptide analogues more fully, we solved the structure of Gla-domainless FVIIa complexed with E-76 by X-ray crystallography (Table 1). With the exception of residues 142–153 (the 140s loop, see below; the chymotrypsinogen numbering system is used for FVIIa), the FVIIa catalytic and EGF-2 domains are almost unchanged from previous FVIIa structures^{20,27,28,45}. The backbone fold within the disulphide-bonded loop of the peptide in the complex is essentially the same

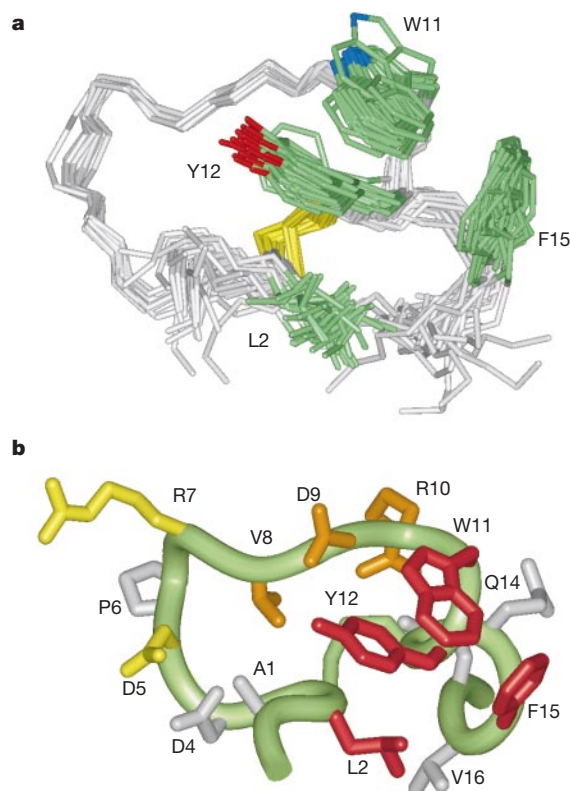


Figure 4 Solution structure of E-76. **a**, Ensemble of E-76 NMR structures. The best 25 structures based on restraint violation energy are shown. Side chains of the cysteines and residues forming the hydrophobic cluster are in yellow and green, respectively. **b**, SAR of E-76 analogues. The fold decreases for the inhibition of FX activation for individual alanine substitutions of E-76 are depicted as red (>1,000), orange (100–1,000), yellow (50–100) and grey (1–50). Similar results were obtained using an FVIIa-binding ELISA (data not shown). Data for the Val 16P analogue refers to a deletion of Val 16P, Glu 17P and Gly 18P. Alanine substitution at the two cysteines resulted in a peptide with >1,000-fold decreased activity and is not shown.

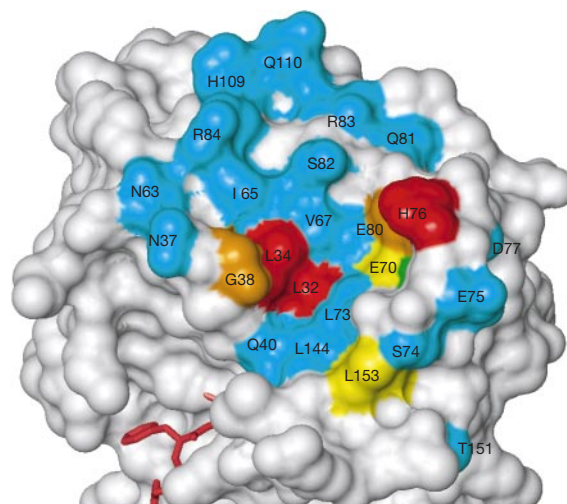


Figure 5 Location of the E-76 binding site. The fold increases in the IC₅₀ for E-76 inhibition of FX activation for individual alanine substitutions of selected surface-exposed residues of FVIIa are depicted as red (>150), orange (~80), yellow (~20) and blue (1). The active site is occupied by D-FFRCMK, shown in red; the Ca²⁺ ion, adjacent to Glu 70 and Glu 80, is shown in green.

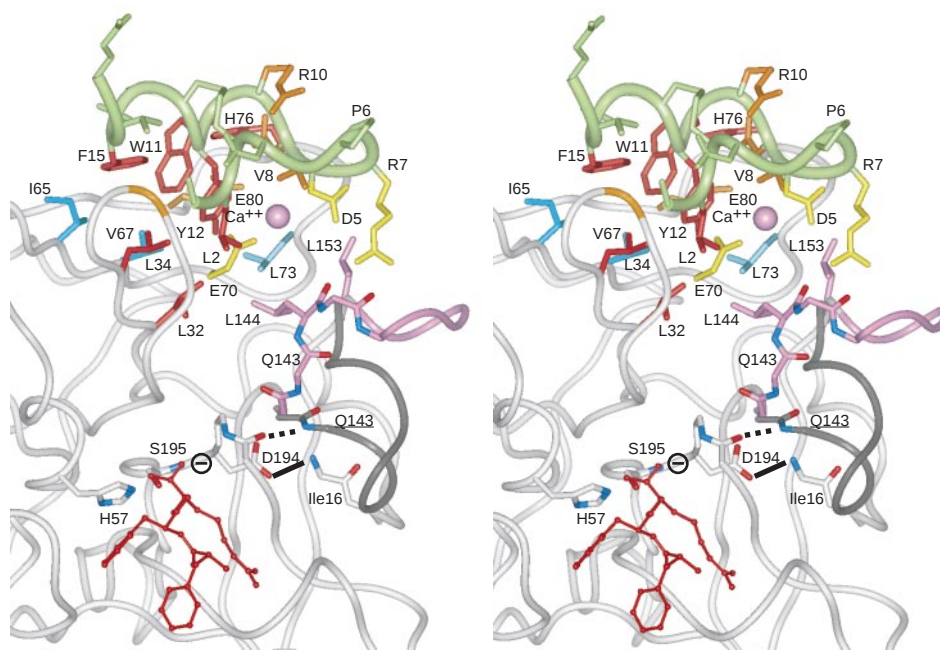


Figure 6 Stereo view of the E-76–FVIIa complex. The backbones of E-76 and FVIIa are represented by green and grey tubes, respectively. Selected side chains in the peptide-binding site are shown and coloured as in Figs 4b and 5, whereas the covalent inhibitor (D-FFRCMK) is shown in red. Other side chains and backbone atoms in the active site or postulated to be important for the mechanism of FX inhibition are also shown (see text). The charge–charge interaction between the N-terminal amine of Ile 16 and the side-chain carboxylate of Asp 194 is shown as a solid line. The conformation of the loop containing residues 142–154 in the presence of E-76 is shown in magenta. The same loop in the absence of E-76 is shown in black²⁰; the hydrogen bond from the amide nitrogen of

Gln 143 (underlined in the structure lacking E-76) to the carbonyl oxygen of Lys 192 is depicted as a dashed line, the amide nitrogen atoms of Gly 193 and Ser 195 that form the oxyanion hole are shown in blue, and the approximate position of the negatively charged oxygen of the tetrahedral intermediate is shown as a dash in a circle. Note that at the current level of resolution, the orientation of the peptide groups forming the oxyanion hole are not absolutely known, although considering the nature of the covalent inhibitor, they are probably oriented as shown. However, the hydrogen bond from the amide nitrogen of Gln 143 to the carbonyl oxygen of Lys 192 is clearly absent in the complex with E-76.

as that observed in solution (r.m.s.d. 1.12 ± 0.06 Å); however, the three aromatic residues on E-76 adopt different side-chain conformations, presumably to maximize interactions with FVIIa.

As expected from the mutagenesis data, E-76 is indeed located at the exosite described in Fig. 5, with its aromatic/hydrophobic face contacting FVIIa and the N-terminus of the helix directed towards the Ca^{2+} -binding loop (Fig. 6). The contact surface on FVIIa is also predominantly hydrophobic, with Leu 32, Leu 34, Ile 65, Val 67 and Leu 73 of the exosite contacting the aromatic side chains of E-76. Intermolecular hydrogen-bonding interactions occur between Tyr 12P and Glu 70 and between Asp 9P and His 76 on the Ca^{2+} -binding loop. There is also a main-chain hydrogen bond from Leu 2P to Gly 38.

Implications for the mechanism of inhibition

The non-competitive nature of the inhibition by E-76 of FX activation by TF–FVIIa (reduced V_{max} , unaltered K_{m} ; Fig. 2) implies that the peptide does not preclude FX from binding to the TF–FVIIa complex. We propose that E-76 binding inhibits FX activation either by an allosteric effect transmitted to the FVIIa active site or by a localized steric effect on FX that would not manifest a change in K_{m} owing to its putative large binding interface. The former possibility is supported by the inhibition by E-76 of small chromogenic substrate turnover (Fig. 1), as E-76 should not block entry of these molecules into the active site (Fig. 6); the incomplete inhibition found is consistent with such an allosteric effect. Alternatively, a localized steric effect on FX by E-76 is plausible as the peptide could interfere with FX binding in a subtle manner that would prevent the scissile bond from obtaining the optimal geometry for turnover. These two hypotheses are not mutually exclusive.

Inspection of the E-76–FVIIa structure suggests how these two

effects on FX activation can be combined into a single mechanism of inhibition by E-76. The most striking difference between FVIIa structures solved previously and that in the current study is a large conformational change (up to 12 Å) in the 140s loop (Fig. 6). Residues in this loop are likely to be involved in FX binding as mutations here affect activation of FX^{17,18}. An important consequence of the loop movement is the loss of a hydrogen bond from the Gln 143 main-chain amide to the carbonyl oxygen of Lys 192. This is significant as the adjacent amide nitrogen of Gly 193 and the amide nitrogen of Ser 195 form the ‘oxyanion hole’, which plays a key catalytic role by stabilizing the negatively charged tetrahedral intermediate (Fig. 6)^{29–31}. Thus, the change in loop conformation induced by E-76 binding at the exosite may inhibit FX activation by both steric (altering the manner in which substrate FX binds) and allosteric (disruption of the oxyanion hole in the enzyme active site) components. This inhibition mechanism is best described as an allosteric ‘switch’ involving the 140s activation loop.

The manner by which E-76 inhibits FX activation is distinct from that of hirugen, a dodecapeptide inhibitor of thrombin derived from the C terminus of the leech protein hirudin, despite topologically similar binding sites in the two proteases^{32,33}. Hirugen binding at the ‘fibrinogen-binding exosite’ of thrombin does not result in a well defined 140s loop, nor does it cause a change in the hydrogen bond from residue 143 to residue 192 or the oxyanion hole. Likewise, the 142-residue thrombin exosite inhibitor triabin does not induce any change in the hydrogen-bond network at the oxyanion hole³⁴. Furthermore, hirugen is a competitive inhibitor of fibrinogen cleavage and does not inhibit the amidolytic activity of small molecule substrates³⁵.

The precise cause of the marked conformational change in the 140s loop of FVIIa in complex with E-76 is not certain. The side

chain of Arg 7P forms two hydrogen bonds with FVIIa main-chain carbonyl oxygens at Leu 144 and Leu 145 (Fig. 6). Additionally, the aliphatic part of the Arg 7P side chain contacts Leu 153, consistent with the 65- and 16-fold increased IC₅₀ values for inhibition of FX activation for alanine substitutions at Arg 7P and Leu 153, respectively. An interaction between the side chains of Leu 2P and Leu 144 is difficult to assess directly, as alanine substitution at Leu 144 leads to a relatively inactive enzyme³⁶ and Leu 2P participates in several other interactions. Finally, the side chains of Leu 2P, Val 8P and the aliphatic part of Arg 7P contact a collection of hydrophobic protein residues including Leu 32, Leu 34, Val 67, Leu 73 and Leu 153. In the FVIIa–E-76 complex, these residues constitute a mini-hydrophobic core. Conceivably, the presence of E-76 is sufficient to nucleate a ‘hydrophobic collapse’ that includes Leu 144 and brings about the change in conformation of the activation loop.

It is interesting to place this large conformational change in the context of zymogen activation in the FVII system. Compared to trypsin and chymotrypsin, the activation process for FVIIa is both more complex and less well characterized^{29–31}. Similar to these related serine proteases, proteolytic cleavage of FVII creates a mature N-terminal residue that eventually becomes buried under the 140s loop (Fig. 6)²⁰. For trypsin and chymotrypsin, the placement of this new N terminus and structural changes in three nearby loops (collectively called the ‘activation domain’) give rise to the active enzyme^{37,38}. For FVII, however, this cleavage and any associated rearrangements are insufficient to form a catalytically competent enzyme; full activity arises only after association with its co-factor, TF. However, although the absence of TF substantially reduces the specific activity of FVIIa against both FX and chromogenic substrates, there was no change in the binding of E-76. The structure of active-site-inhibited Gla-domainless FVIIa in the absence of TF has shown that the changes caused by TF binding do not extend to large changes in the 140s loop, at least in the context of an occupied active site²⁸. Residues 165–170 in our structure are in the same conformation as that of ref. 20 and are not in accord with the proposal that changes observed here are due to the absence of TF²⁸. When we consider the possibility that the loops of the activation domain adopt distinct structures in each of the (at least) three biochemically distinct forms (zymogen FVII, FVIIa and TF–FVIIa), it is reasonable to speculate that the 140s loop in our structure bears a close relationship to that in zymogen FVII. Thus E-76 may induce a zymogen or zymogen-like structure, resulting in the observed inhibition.

Conclusions

We have shown that a structured peptide selected from a naive library binds with high affinity and selectivity to an exosite on FVIIa. Unexpectedly, this peptide also potentially inhibits the TF–FVIIa-catalysed activation of FX using a non-competitive type of allosteric mechanism and is therefore a potent anticoagulant as well. This provides an opportunity to develop anticoagulants based upon an entirely new site, distinct from the active site. The success of this approach in finding potent exosite inhibitors for FVIIa represents a new framework for developing inhibitors of serine proteases. □

Methods

Selection of phage-displayed peptides

TF_{1–243} or FVIIa (2 µg ml^{−1} each) in 50 mM sodium carbonate (pH 9.3) were immobilized to Maxisorp plates (Nunc). Wells were blocked using sorting buffer (50 mM Hepes (pH 7.2), 5 mM CaCl₂, 5 mM MgCl₂, 150 mM NaCl, 1% BSA). To form TF–FVIIa, we added FVIIa (2 µg ml^{−1}) in sorting buffer to blocked TF-coated wells. Three pools of polyvalent naive 20-residue peptide phage libraries²³ (pool A (X_{6–7}CX_{5–6}), pool B (X₄CX₂GPX₄CX₄, X₂₀ and X₇CX₄CX₇) and pool E (X_{4–5}CX_{8–10}CX_{4–5})), each containing ~5 × 10¹⁰ phage, were separately incubated with immobilized targets in sorting buffer for 3 h at 25 °C. Wells were washed with 50 mM Hepes (pH 7.2), 150 mM NaCl, 0.005% Tween 20; phage were eluted with 500 mM KCl, 10 mM HCl, pH 2 and amplified as described²³.

The E-56 sequence (EAAALCDPRLDWYCIAGE) was matured using monovalent phage display²⁴. We designed partially randomized libraries to maintain a bias towards E-56 using oligonucleotides with a 70–10–10–10 mixture of bases³⁹. Completely random-

ized libraries used oligonucleotides containing NNS for particular codons (N = G/A/T/C; S = G/C). Peptides were synthesized as described²³.

Peptide binding and inhibition

We measured inhibition of FX activation (Fig. 1) with 300 pM relipidated TF_{1–243}, 20 pM FVIIa and 165 nM FX at 25 °C as described⁴⁰. In Fig. 2a we used 160 pM relipidated TF_{1–243} and 16 pM FVIIa in assay buffer (100 mM Hepes (pH 7.8), 140 mM NaCl, 5 mM CaCl₂, 0.02% Tween 80) containing 0.5% BSA. Amidolytic activity was measured at 405 nm using 10 nM FVIIa, 60 nM TF_{1–219} and 1.33 mM Chromozym t-PA in assay buffer containing 0.1% PEG 8000 at 25 °C. Amidolytic assays without TF_{1–219} used 3.33 mM substrate; activities were 6.6-fold lower, in accord with previous results²⁶.

We determined binding constants using a BIAcore 2000. Following immobilization of C-terminal biotinylated E-76 to a streptavidin chip, we monitored binding and dissociation of FVIIa with or without 10 µM TF_{1–219} in 20 mM Tris, pH 7.5, 100 mM NaCl, 5 mM CaCl₂ and 0.01% Tween 20.

NMR spectroscopy

One- and two-dimensional (2QF-COSY, TOCSY, ROESY, NOESY) NMR spectra of E-76 were acquired on a Bruker AMX500 spectrometer at 25 °C as described²³; samples contained a 2.5 mM solution of peptide at pH 6.1. Structural restraints (118 interproton distances, 12 φ dihedral angle restraints and 4 χ₁ dihedral angle restraints) were obtained as described⁴¹. Explicit hydrogen bond restraints were not employed. Structures were calculated using DGII (CVFF force field parameters) and refined using DISCOVER (AMBER all atom force field parameters) (MSI)⁴¹. The 25 conformations of lowest restraint violation energy (0.03 to 0.26 kcal mol^{−1}) contained no violation of distance and dihedral angle restraints greater than 0.07 Å and 1.9°, respectively. The mean r.m.s.d. to the mean structure for N, C^α and C atoms of residues Cys 3P to Cys 13P was 0.30 ± 0.09 Å.

Peptide-binding exosite on FVIIa

Plasmids encoding FVII/FIX chimaeras VII-EGF2/IX (FVII/FIX_{protease}) and IX-EGF2/VII (FVII_{protease}/FIX) were transiently transfected in 293 cells (ATCC CRL 1573)⁴². Proteins were captured using TF_{1–219} or an antibody against FVII. Specific binding of phage displaying E-76 was detected using HRP-conjugated anti-M13 (Amersham Pharmacia Biotech).

Substitution of rabbit FVIIa (GenBank accession no. U77477; B. J. Clarke, personal communication) sequences or alanine residues into human FVIIa by site-directed mutagenesis⁴³ was confirmed by DNA sequencing. FVII expression in 293 cells was quantified by ELISA; western blot analysis and E-76 phage binding were also determined. Specific activities for FX activation by FVII mutants and subsequent inhibition by E-76 were determined essentially as described above.

FVIIa purification and X-ray crystallography with E-76

Cell-culture medium from 293 cells expressing human factor VII brought to 5 mM EDTA/2 mM benzamidine was loaded onto a DEAE column and eluted with 150 mM NaCl. This pool was loaded onto a Q-Sepharose column and eluted with 5 mM CaCl₂, 135 mM NaCl. The Q-Sepharose pool eluted as activated FVIIa and was precipitated with ammonium sulphate. Solubilized FVIIa was passed over a Sephacryl-200 column; the N-terminal residue was Ile 42, confirming the loss of the Gla domain. E-76 and Gla-domainless FVIIa, inhibited with D-Phe-L-Phe-Arg-chloromethyl ketone (D-FRCKMK), in a 1:1 ratio, were co-purified on a Superdex 75 column. Crystals from hanging drops equilibrated against 20% (w/v) PEG 4000, 10% *t*-butanol and 0.1 M sodium cacodylate (pH 5.5) were harvested into cryoprotectant-containing reservoir plus 10% methylpentanediol and flash frozen in liquid nitrogen. Data to 3 Å from a MAR345 scanner (space group P2₁ with *a* = 70.49 Å, *b* = 55.26 Å, *c* = 111.73 Å, β = 99.48°) at SSRL beamline 9.1 were reduced with Mosflm and CCP4 (*R*_{merge} = 10.4% (26.2% outer shell); *I*/σ = 6.2 (2.7); completeness 98% (99%)). Solution by molecular replacement (Amore) used the TF–FVIIa structure²⁰. Refinement and model adjustment used X-PLOR v3.851 and Xsight (MSI). We applied non-crystallographic symmetry restraints (two complexes per asymmetric unit (r.m.s.d. = 0.44 Å)). The model includes residues 16–257 of the catalytic domains, residues 42–142 of the light chains, residues 1–17 of the E-76 peptides, two active-site inhibitors, 4 Ca²⁺ ions, 2 cacodylate ions, 5 sugar residues and 4 waters.

Received 18 November 1999; accepted 17 February 2000.

- Nemerson, Y. Tissue factor and hemostasis. *Blood* **71**, 1–8 (1988).
- Rapaport, S. I. & Rao, L. V. M. The tissue factor pathway: How it has become a ‘prima ballerina’. *Thromb. Haemost.* **74**, 7–17 (1995).
- Davie, E. W., Fujikawa, K. & Kisiel, W. The coagulation cascade: Initiation, maintenance, and regulation. *Biochemistry* **30**, 10363–10370 (1991).
- Hirsh, J. & Weitz, J. I. Thrombosis: New antithrombotic agents. *Lancet* **353**, 1431–1436 (1999).
- Vlasuk, G. P. The new anticoagulants: New opportunities, new issues. *Arch. Pathol. Lab. Med.* **122**, 812–814 (1998).
- Hirsh, J., Ginsberg, J. S. & Marder, V. J. In *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (eds Colman, R. W., Hirsh, J., Marder, V. J. & Salzman, E. W.) 1567–1583 (J. B. Lippincott, Philadelphia, 1994).
- Weitz, J. I. Low molecular weight heparins. *N. Engl. J. Med.* **337**, 688–698 (1997).
- Hirsh, J. Heparin. *N. Engl. J. Med.* **324**, 1565–1574 (1991).
- Broze, G. J. Jr, Girard, T. J. & Novotny, W. F. Regulation of coagulation by a multivalent Kunitz-type inhibitor. *Biochemistry* **29**, 7539–7546 (1990).
- Broze, G. J. Jr, Likert, K. & Higuchi, D. Inhibition of Factor VIIa/tissue factor by antithrombin III and tissue factor pathway inhibitor. *Blood* **82**, 1679–1680 (1993).

11. Mann, K. G. Inhibition of Factor VIIa/tissue factor by antithrombin III and tissue factor pathway inhibitor. *Blood* **82**, 1680–1681 (1993).
12. Bode, W. & Huber, R. Natural protein protease inhibitors and their interaction with proteinases. *Eur. J. Biochem.* **204**, 433–451 (1992).
13. Laskowski, M. Jr & Kato, I. Protein inhibitors of proteinases. *Annu. Rev. Biochem.* **49**, 593–626 (1980).
14. Sanderson, P. E. J. Small, noncovalent serine protease inhibitors. *Med. Res. Rev.* **19**, 179–197 (1999).
15. Stubbs, M. T. & Bode, W. Coagulation factors and their inhibitors. *Curr. Opin. Struct. Biol.* **4**, 823–832 (1994).
16. Bode, W., Brandstetter, H., Mather, T. & Stubbs, M. T. Comparative analysis of haemostatic proteinases: Structural aspects of thrombin, Factor Xa, Factor IXa, and Protein C. *Thromb. Haemost.* **78**, 501–511 (1997).
17. Kirchhofer, D. & Banner, D. W. Molecular and structural advances in tissue factor-dependent coagulation. *Trends Cardiovasc. Med.* **7**, 316–324 (1997).
18. Ruf, W. & Dickinson, C. D. Allosteric regulation of the cofactor-dependent serine protease coagulation Factor VIIa. *Trends Cardiovasc. Med.* **8**, 350–356 (1998).
19. Higashi, S. & Iwanaga, S. Molecular interaction between factor VII and tissue factor. *Int. J. Hematol.* **67**, 229–241 (1998).
20. Banner, D. W. *et al.* The crystal structure of the complex of blood coagulation factor VIIa with soluble tissue factor. *Nature* **380**, 41–46 (1996).
21. Johnson, K. & Hung, D. Novel anticoagulants based on inhibition of the factor VIIa/tissue factor pathway. *Coron. Artery Dis.* **9**, 83–87 (1998).
22. Lowman, H. B. Bacteriophage display and discovery of peptide leads for drug development. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 401–424 (1997).
23. Lowman, H. B. *et al.* Molecular mimics of insulin-like growth factor 1 (IGF-1) for inhibiting IGF-1: IGF-binding protein interactions. *Biochemistry* **37**, 8870–8878 (1998).
24. Dennis, M. S. & Lazarus, R. A. Kunitz domain inhibitors of tissue factor-Factor VIIa I. Potent inhibitors selected from libraries by phage display. *J. Biol. Chem.* **269**, 22129–22136 (1994).
25. Segel, I. H. in *Enzyme Kinetics* 100–226 (Wiley, New York, 1975).
26. Neuenchwander, P. F., Branam, D. E. & Morrissey, J. H. Importance of substrate composition, pH, and other variables on tissue factor enhancement of Factor VIIa activity. *Thromb. Haemost.* **70**, 970–977 (1993).
27. Zhang, E., St. Charles, R. & Tulinsky, A. Structure of extracellular tissue factor complexed with Factor VIIa inhibited with a BPTI mutant. *J. Mol. Biol.* **285**, 2089–2104 (1999).
28. Pike, A. C. W., Brzozowski, A. M., Roberts, S. M., Olsen, O. H. & Persson, E. Structure of human factor VIIa and its implications for the triggering of blood coagulation. *Proc. Natl Acad. Sci. USA* **96**, 8925–8930 (1999).
29. Kraut, J. Serine proteases: Structure and mechanism of catalysis. *Annu. Rev. Biochem.* **46**, 331–358 (1977).
30. Steitz, T. A. & Shulman, R. G. Crystallographic and NMR studies of the serine proteases. *Annu. Rev. Biophys. Bioeng.* **11**, 419–444 (1982).
31. Polgár, L. in *Mechanisms of Protease Action* 87–122 (CRC, Boca Raton, 1989).
32. Skrzypczak-Jankun, E. *et al.* Structure of hirugen and hirulog 1 complexes of alpha-thrombin. *J. Mol. Biol.* **221**, 1379–1393 (1991).
33. Stubbs, M. T. & Bode, W. A player of many parts: The spotlight falls on thrombin's structure. *Thromb. Res.* **69**, 1–58 (1993).
34. Fuentes-Prior, P. *et al.* Structure of the thrombin complex with triabin, a lipocalin-like exosite-binding inhibitor derived from a triatomine bug. *Proc. Natl Acad. Sci. USA* **94**, 11845–11850 (1997).
35. Naski, M. C., Fenton, J. W. L., Maraganore, J. M., Olson, S. T. & Shafer, J. A. The COOH-terminal domain of hirudin. An exosite-directed competitive inhibitor of the action of alpha-thrombin on fibrinogen. *J. Biol. Chem.* **265**, 13484–13489 (1990).
36. Shobe, J., Dickinson, C. D. & Ruf, W. Regulation of the catalytic function of coagulation Factor VIIa by a conformational linkage of surface residue Glu 154 to the active site. *Biochemistry* **38**, 2745–2751 (1999).
37. Huber, R. & Bode, W. Structural basis of the activation and action of trypsin. *Acc. Chem. Res.* **11**, 114–121 (1978).
38. Bode, W. & Huber, R. in *Molecular and Cellular Basis of Digestion* (eds Desnuelle, P., Sjöström, H. & Norén, O.) 213–234 (Elsevier, Amsterdam, 1986).
39. Gallop, M. A., Barrett, R. W., Dower, W. J., Fodor, S. P. A. & Gordon, E. M. Applications of combinatorial technologies to drug discovery. *J. Med. Chem.* **37**, 1233–1251 (1994).
40. Kelley, R. F. *et al.* A soluble tissue factor mutant is a selective anticoagulant and antithrombotic agent. *Blood* **89**, 3219–3227 (1997).
41. Skelton, N. J., Garcia, K. C., Goeddel, D. V., Quan, C. & Burnier, J. P. Determination of the solution structure of guanylin. *Biochemistry* **33**, 13581–13592 (1994).
42. Toomey, J. R., Smith, K. J. & Stafford, D. W. Localization of the human tissue factor recognition determinant of human Factor VIIa. *J. Biol. Chem.* **266**, 19198–19202 (1991).
43. Kunkel, T. A., Roberts, J. D. & Zakour, R. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol.* **154**, 367–382 (1987).
44. Marquardt, D. W. An algorithm for least squares estimation of non linear parameters. *J. Soc. Indust. Appl. Math.* **11**, 431–441 (1963).
45. Kembball-Cook, G., Johnson, D. J. D., Tuddenham, E. G. D. & Harlos, K. Crystal structure of active site-inhibited human coagulation factor VIIa (des-Glu). *J. Struct. Biol.* **127**, 213–223 (1999).

Acknowledgements

We thank R. Kelley for TF and helpful discussions; R. Artis, R. McDowell, C. Refino and D. Kirchhofer for helpful discussions; R. Smyth and S. Bullens for clotting assays; H. Lowman, B. Cunningham and G. Nakamura for help with library constructions; C. Quan, J. Tom and M. Struble for peptide synthesis; M. Beresini, A. Hebert and L. Caris for ELISA assays; W. Prince for inhibition assays; P. Jhurani, P. Ng and M. Vasser for DNA synthesis; M. Hamner, A. Zhong and A. Goddard for DNA sequencing; D. Stafford and J. Toomey for plasmids encoding FVII and FVII/FIX chimaeras; B. J. Clarke for a preprint on rabbit FVIIa expression and characterization; and A. de Vos, J. Burnier and D. Lowe for their support. This work is based in part upon research conducted at the Stanford Synchrotron Radiation Laboratory (SSRL), which is funded by the Department of Energy, Office of Basic Energy Sciences.

Correspondence and requests for materials should be addressed to R.A.L. (e-mail: lazarus.bob@gene.com). The X-ray crystallographic coordinates of FVIIa complexed with E-76 have been deposited in the Protein Data Bank (accession code 1DVA).

Double-quantum vortex in superfluid $^3\text{He-A}$

R. Blaauwgeers^{*†}, V. B. Eltsov^{*‡}, M. Krusius^{*}, J. J. Ruohio^{*}, R. Schanen^{*§} & G. E. Volovik^{*||}

^{*} Low Temperature Laboratory, Helsinki University of Technology, PO Box 2200, FIN-02015 HUT, Finland

[†] Kamerlingh Onnes Laboratory, Leiden University, PO Box 9504, 2300 RA Leiden, The Netherlands

[‡] Kapitza Institute for Physical Problems, Kosygina 2, 117334 Moscow, Russia

[§] Centre de Recherches sur les Très Basses Températures, Centre National de la Recherche Scientifique, BP 166, F-38042 Grenoble, Cedex 09, France

^{||} Landau Institute for Theoretical Physics, Kosygina 2, 117334 Moscow, Russia

Linear defects are generic in continuous media¹. In quantum systems they appear as topological line defects which are associated with a circulating persistent current. In relativistic quantum field theories they are known as cosmic strings², in superconductors as quantized flux lines³, and in superfluids^{3,4} and low-density Bose–Einstein condensates⁵ as quantized vortex lines. A conventional quantized vortex line consists of a central core around which the phase of the order parameter winds by $2\pi n$, while within the core the order parameter vanishes or is depleted from the bulk value. Usually vortices are singly quantized (that is, have $n = 1$). But it has been theoretically predicted that, in superfluid $^3\text{He-A}$, vortex lines are possible that have $n = 2$ and continuous structure, so that the orientation of the multi-component order parameter changes smoothly throughout the vortex while the amplitude remains constant. Here we report direct proof, based on high-resolution nuclear magnetic resonance measurements, that the most common vortex line in $^3\text{He-A}$ has $n = 2$. One vortex line after another is observed to form in a regular periodic process, similar to a phase-slip in the Josephson effect.

According to the original Landau picture of superfluidity, superflow is potential: its velocity $\mathbf{v}_s$ is curl-free ($\nabla \times \mathbf{v}_s = 0$). Later Onsager⁶ and Feynman⁷ found that this statement must be generalized: $\nabla \times \mathbf{v}_s$ can be non-zero at a singular line, the core of a quantized vortex line, around which the phase of the order parameter winds by $2\pi n$. The discovery of superfluid $^3\text{He-A}$, with Cooper pairing in an orbital $L = 1$ and spin $S = 1$ state, introduced further reservations to Landau's rule: nonsingular superfluid vorticity can also be produced by a regular order-parameter texture, which consists of the spatial variation in the orientation of the orbital quantization axis⁸. It has recently been suggested that vorticity with continuous structure may also exist in spinor states of Bose gases with a vector Bose condensate⁹. In some relativistic quantum field models, further analogues appear in the form of continuous strings or stringy textures¹⁰.

In $^3\text{He-A}$, the local orientation of the orbital quantization axis is denoted with the unit vector $\hat{\mathbf{l}}$ (the analogue of isospin in the standard model of particle physics¹⁰). It points in the direction of the orbital momenta of the Cooper pairs. Quantization of vorticity now becomes possible through the topology of a continuously distributed $\hat{\mathbf{l}}$ field or texture: a structure, where the $\hat{\mathbf{l}}$ orientations cover all 4π spatial directions, produces around itself a winding of the order parameter phase by 4π (ref. 11). In a rotating container such structures are stacked like vortex lines (Fig. 1) so that the $\hat{\mathbf{l}}$ texture is translationally invariant in the direction of the rotation axis, but looks like a periodic lattice in the transverse plane¹².

The first indications for the existence of continuous vorticity were seen in rotating NMR experiments¹³ on $^3\text{He-A}$ which revealed a small satellite peak in the NMR spectrum during rotation (Fig. 1, bottom). The frequency shift and size of this NMR line were more

consistent with calculations based on a vortex structure with continuous $\hat{\mathbf{l}}$ -field distribution rather than one with a singular hard vortex core and only 2π phase winding¹⁴.

In our experiment the liquid $^3\text{He-A}$ sample is contained in a cylinder with a diameter $2R = 3.87$ mm and a height of 8 mm. It is rotated by rotating the entire nuclear demagnetization refrigerator. The NMR measurements are performed in a magnetic field of 9 mT. The sharpest signals of single-vortex formation are recorded when the field is oriented perpendicular to the rotation axis. In Fig. 2 the height of the vortex satellite peak has been plotted as a function of the rotation drive Ω . During sufficiently slow acceleration ($|\dot{\Omega}| < 5 \times 10^{-4} \text{ rad s}^{-2}$) a staircase-like response is revealed: vortex lines are formed regularly in a periodic process one by one, at a stable constant value of critical velocity, each time when the rotation has been increased by a constant increment $\Delta\Omega$.

A compact interpretation¹¹ of these macroscopic phase-slip events can be presented in terms of the Josephson effect, as was directly verified by measurements on the precession time of a vortex filament in $^3\text{He-B}$ (ref. 15). The frequency f of $2\pi n$ phase-slip events in pair-correlated superfluids and superconductors obeys the Josephson–Anderson relation

$$f = (2/n)(\Delta\mu/h) \quad (1)$$

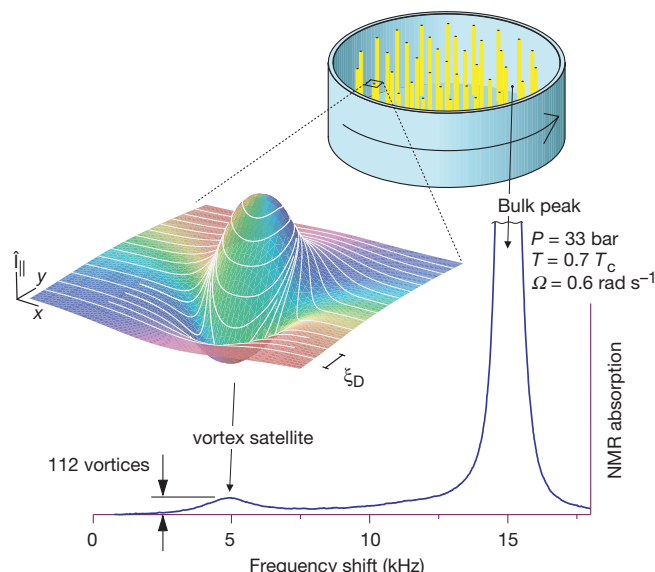


Figure 1 Double-quantum vortex in superfluid $^3\text{He-A}$ and its nuclear magnetic resonance signature. Top: in macroscopically coherent systems, such as superfluids or type II superconductors, linear quantized defects are formed under the influence of an appropriate external field. When a cylindrical container with superfluid is rotated around its symmetry axis, quantized vortex lines will start to form above a certain critical angular rotation velocity. Centre: the orbital $\hat{\mathbf{l}}$ field of a 4π vortex in a magnetic field consists of a soft vortex core and a surrounding homogeneous texture. Within the soft core the order-parameter amplitude remains constant, but the orientation of $\hat{\mathbf{l}}$ varies continuously over all spatial directions. The magnitude of the axial component $\hat{l}_{\parallel}$ (parallel to Ω) is plotted on the vertical axis. The colours denote the order parameter phase (blue, 0 and 2π ; red, π). Half of the vorticity within the soft core is concentrated in the hump ($\hat{l}_{\parallel} > 0$), while the second half is located in the region of the pit ($\hat{l}_{\parallel} < 0$). The orientation of the transverse component of $\hat{\mathbf{l}}$ is tangential to the superimposed white lines. The length scale for the soft vortex core is the healing length of the spin–orbit interaction $\xi_D \approx 10 \mu\text{m}$. Bottom: the spin–orbit interaction provides an extra torque on NMR precession, so that the resulting frequency shift depends on the relative orientation of the spin and orbital momenta of the Cooper pair. Inside the soft core the $\hat{\mathbf{l}}$ orientation changes continuously. This gives rise to absorption in a small characteristic satellite peak. Its frequency is shifted from the large bulk liquid peak, which in turn originates from the homogeneously aligned $\hat{\mathbf{l}}$ field outside the soft core. Thus the soft cores are isolated from each other and contribute equally to the satellite absorption, which provides a practical tool for vortex monitoring.

Here $\Delta\mu$ is the effective difference in the chemical potential along the boundary of the container, as produced by the increase in Ω : $\Delta\mu = m_3 \cdot R_{\text{eff}} \cdot \Omega \cdot 2\pi R_{\text{eff}}$, where m_3 is the mass of the ^3He atom and R_{eff} ($\approx R$) the distance of the phase-slip site from the centre of the container. The increment in Ω between two successive phase-slip events is $\Delta\Omega = \dot{\Omega}/f$ or

$$\Delta\Omega = \frac{nh}{4\pi m_3 R_{\text{eff}}^2} \quad (2)$$

Thus the winding number n can be extracted from a measurement of the increment $\Delta\Omega$. The minimum value of $\Delta\Omega$ for a vortex with $n = 2$ is $5.6 \times 10^{-3} \text{ rad s}^{-1}$, when the vortex is formed at the wall of the container ($R_{\text{eff}} = R$). The Josephson–Anderson frequency is thus expected to be $f \approx 0.05 \text{ Hz}$, which is comparable to earlier experiments^{15,16} on superfluid ^4He and $^3\text{He-B}$.

In Fig. 3, the distribution function of the measured $\Delta\Omega$ values is plotted. It has a maximum at $\Delta\Omega_m = 6.2 \times 10^{-3} \text{ rad s}^{-1}$, which is close to the expected value, and falls down smoothly at $\Delta\Omega_m/2$ and $3\Delta\Omega_m/2$. This suggests that the Josephson–Anderson relation is satisfied with $n = 2$ and that the phase slip is associated with the creation of a 4π vortex. Although there are no $n = 1$ or $n = 3$ events within experimental precision, the distribution function is not entirely symmetric, owing to one measuring run in which $n = 4$ events occur periodically, as shown by the uppermost signal trace in Fig. 2. A similar exceptional example of two vortex lines appearing simultaneously was observed¹⁶ in $^3\text{He-B}$.

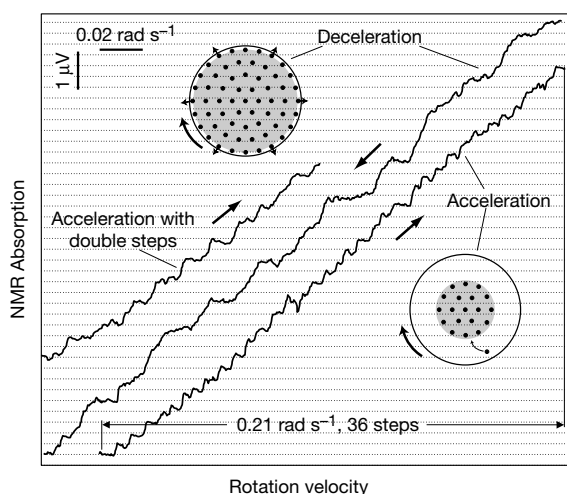


Figure 2 Height of the vortex satellite peak in the NMR absorption spectrum, as a function of rotation velocity Ω . The lowest and uppermost traces show the formation of vortex lines during acceleration, and the trace in the middle corresponds to the annihilation of vortex lines during deceleration. The traces exhibit a 'staircase' pattern, which arises from the discrete nature of the quantized superfluid vorticity. All traces are plotted with the same horizontal and vertical scales, but have been shifted for easier comparison. The distance between the horizontal grid lines corresponds to the expected absorption change per one $n = 2$ vortex, derived independently from the change in the peak height of the satellite over a large difference in Ω . We note the different patterns during rotational acceleration and deceleration: vortex creation takes place in the form of a regular periodic process with one vortex line forming after another close to the container wall and then moving across the vortex-free region to the vortex cluster (see cross-section of container at bottom right). In contrast, the annihilation of vortex lines is a more random process, with larger steps and plateaux: here several lines in the outermost circle may be pushed to the cylindrical wall roughly simultaneously by random irregularities in rotation, owing to misalignment of container and rotation axes, or asymmetry and faults in the order-parameter texture (see cross-section at upper left). The uppermost trace shows a rare case of rotational acceleration where a pair of $n = 2$ vortex lines is simultaneously formed in a periodic process.

The staircase response in Fig. 2 is blurred by noise in the measurement and by relaxation effects in the vortex cluster (which occur after the addition of each new line). Nevertheless, the correlation analysis in Fig. 4 confirms the presence of a coherent long-range periodicity in the signal amplitude $A(\Omega)$ from the phase-slip process: The correlation function $\langle A(\Omega)A(\Omega - \Delta\Omega) \rangle$

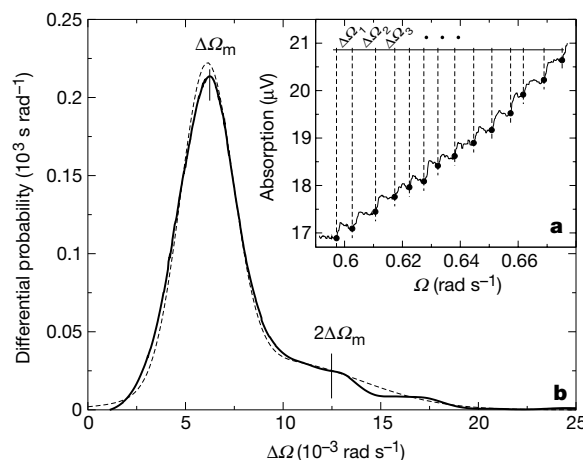


Figure 3 Determination of the probability distribution for the interval $\Delta\Omega$ which separates successive phase-slip events in periodic vortex formation. **a**, The example shows how individual $\Delta\Omega$ values are extracted from the measured staircase records. The black dots are entered on the record by eye and the corresponding $\Delta\Omega$ values are read out. From more than 200 analysed steps the integral probability distribution of step sizes is worked out. It is then smoothed and differentiated, to yield the differential probability distribution in **b** (solid line). It peaks at $\Delta\Omega_m = (6.24 \pm 0.13) \times 10^{-3} \text{ rad s}^{-1}$. According to equation (2), this corresponds to $n = 2$ vortices which are created close to the wall of the cylinder. At $2\Delta\Omega_m$ a small contribution from two-vortex events is visible. The entire distribution function can be fitted with two gaussian lines which account for $n = 2$ and $n = 4$ events (dashed line in **b**).

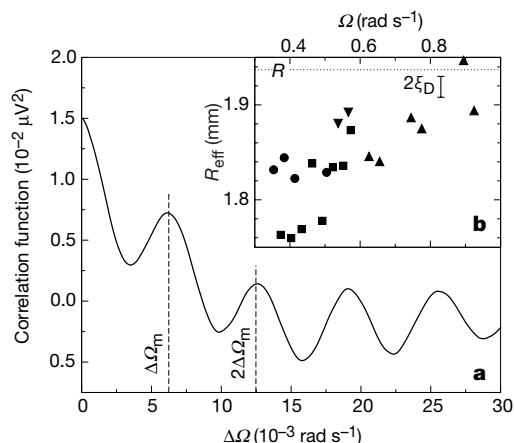


Figure 4 Stable long-range periodicity of phase-slip process. **a**, From the record of absorption amplitude A as a function of Ω in Fig. 2, the correlation function $F(\Delta\Omega) = (\Omega_2 - \Omega_1)^{-1} \int_{\Omega_1}^{\Omega_2} A(\Omega')A(\Omega' - \Delta\Omega)d\Omega'$ has been calculated in the interval between $\Omega_1 = 0.48 \text{ rad s}^{-1}$ and $\Omega_2 = 0.59 \text{ rad s}^{-1}$. The periodic component in $A(\Omega)$ transforms to a set of evenly spaced peaks, while noise transforms to a slowly varying background. The positions of the peaks are located at the multiples of the period $\Delta\Omega_m$ of the vortex formation process. The value of $\Delta\Omega_m$ agrees with that which was obtained from the probability distribution in Fig. 3, that is, with $n = 2$ vortices. **b**, The correlation analysis shows that the periodicity in vortex formation is weakly rotation velocity dependent. This translates to the Ω dependence of the radial position R_{eff} of the phase-slip site according to equation (2). In this plot $\Omega_2 - \Omega_1 = 0.06 \text{ rad s}^{-1}$ and the different symbols of the data points denote different runs of rotational acceleration.

has a sinusoidal appearance, with multiple maxima repeating at regular harmonics of $\Delta\Omega_m$. (In the case of an ideal staircase record the sinusoidal curve would be replaced by a succession of evenly spaced sharp peaks.) Here $\Delta\Omega_m$ is determined more precisely and information can be extracted on the radial location $R_{\text{eff}}(\Omega)$ of the phase-slip site, to gain insight into the vortex formation process.

In $^3\text{He-A}$, both $n = 1$ and $n = 2$ vortices exist. In our experimental conditions $n = 1$ vortices have the lowest energy¹⁷. However, it is not energy considerations but the critical properties which matter in the phase-slip process. An emerging new vortex is formed as an elementary vortex ring, comparable in size to the length scale which characterizes its structure. The appropriate length scale for the singular core of the $n = 1$ vortex is the superfluid coherence length ($\xi \approx 10\text{--}100\text{ nm}$), while for the $n = 2$ vortex it is the healing length of the $\hat{\mathbf{l}}$ texture ($\xi_D \approx 10\text{ }\mu\text{m}$). The critical flow velocity v_c for creating the elementary vortex ring is inversely proportional to its size¹⁶ and is thus generally an order of magnitude larger for the $n = 1$ vortex. In this case it also depends on the surface roughness of the cylinder wall as the flow velocity is expected to reach its absolute maximum value in the vicinity of a sharp surface asperity (of approximate size ξ). In contrast, the soft-core of the $n = 2$ vortex is created deep in the bulk liquid, as seen in Fig. 4b. The site of this phase slip becomes Ω dependent, because the distance from the wall $R - R_{\text{eff}}$ is a significant fraction of the width $v_c/(2\Omega)$ of the vortex-free region around the vortex cluster (Fig. 2): when the width shrinks with increasing Ω the phase-slip centre is pushed closer to the wall.

Thus, when superfluid $^3\text{He-A}$ is slowly set into rotation, doubly quantized vortex lines are formed first and so the flow velocity never reaches high enough values for singly quantized events to become possible. This process can be understood as a macroscopic phase-slip phenomenon, as the full length of the vortex line is $\sim 1\text{ cm}$ and the Josephson period is larger than 1 s . The two circulation quanta of the vortex line correspond to a 4π phase winding around the vortex, which arises from a continuous orientational distribution over 4π of the orbital $\hat{\mathbf{l}}$ field within the vortex. $\square$

Received 30 November 1999; accepted 16 February 2000.

- Kléman, M. *Points, Lines, and Walls in Liquid Crystals, Magnetic Systems and Various Ordered Media* (Wiley, New York, 1983).
- Hindmarsh, M. & Kibble, T. W. B. Cosmic strings. *Rep. Prog. Phys.* **58**, 477–562 (1995).
- Tilley, D. R. & Tilley, J. *Superfluidity and Superconductivity* (IOP, New York, 1990).
- Vollhardt, D. & Wölfle, P. *The Superfluid Phases of ^3He* (Taylor & Francis, London, 1990).
- Matthews, M. R. *et al.* Vortices in a Bose–Einstein condensate. *Phys. Rev. Lett.* **83**, 2498–2501 (1999).
- Onsager, L. *Nuovo Cimento* **6** (Suppl. 2), 249 (1949).
- Feynman, R. P. *Progress in Low Temperature Physics* (ed. Gorter, C. G.) Vol. 1, 17–51 (North-Holland, Amsterdam, 1955).
- Mermin, N. D. & Ho, T. L. Circulation and angular momentum in the A phase of superfluid helium-3. *Phys. Rev. Lett.* **36**, 594–597 (1976).
- Ho, T. L. Spinor Bose condensates in optical traps. *Phys. Rev. Lett.* **81**, 742–745 (1998).
- Achucarro, A. & Vachaspati, T. Semilocal and electroweak strings. *Phys. Rep.* (in the press); preprint hep-ph/9904229 at (xxx.lanl.gov) (1999).
- Anderson, P. W. & Toulouse, G. Phase slippage without vortex cores: Vortex textures in superfluid ^3He . *Phys. Rev. Lett.* **38**, 508–511 (1977).
- Volovik, G. E. & Kopnin, N. B. On rotating $^3\text{He-A}$. *JETP Lett.* **25**, 22–24 (1977).
- Hakonen, P. J., Ikkala, O. T. & Islander, S. T. Experiments on vortices in rotating superfluid $^3\text{He-A}$. *Phys. Rev. Lett.* **49**, 1258–1261 (1982).
- Seppälä, H. K. & Volovik, G. E. Evidence for nonsingular vorticity in the Helsinki experiments on rotating ^3He . *J. Low Temp. Phys.* **51**, 279–290 (1983).
- Packard, R. E. The role of the Josephson–Anderson equation in superfluid helium. *Rev. Mod. Phys.* **70**, 641–651 (1998).
- Ruutu, V. M. H., Parts, Ü., Koivuniemi, J. H., Kopnin, N. B. & Krusius, M. Intrinsic and extrinsic mechanisms of vortex formation in rotating superfluid $^3\text{He-B}$. *J. Low Temp. Phys.* **107**, 93–164 (1997).
- Parts, Ü *et al.* Phase diagram of vortices in superfluid $^3\text{He-A}$. *Phys. Rev. Lett.* **75**, 3320–3323 (1995).

Acknowledgements

We thank H. Götz for help with the data analysis. This collaboration was carried out under the EU-TMR programme.

Correspondence and requests for materials should be addressed to V.B.E. (e-mail: ve@boojum.hut.fi).

Imaging of localized electronic states in the quantum Hall regime

N. B. Zhitenev*, T. A. Fulton*, A. Yacoby*†, H. F. Hess*‡, L. N. Pfeiffer* & K. W. West*

* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

† Weizmann Institute of Science, Rehovot 76100, Israel

‡ Phasemetrics Inc., San Diego, California 92121, USA

The concept of electron localization has long been accepted to be essential to the physics of the quantum Hall effect^{1,2} in a two-dimensional electron gas. The exact quantization of the Hall resistance and the zero of the diagonal resistance over a range of filling factors close to integral are attributed to the localization of electronic states at the Fermi level in the interior of the gas. As the electron density is changed, charging of the individual localized states may occur by single-electron jumps^{3,4}, causing associated oscillations in the local electrostatic potential. Here we search for such a manifestation of localized states in the quantum Hall regime, using a scanning electrometer probe^{5,6}. We observe localized potential signals, at numerous locations, that oscillate with changing electron density. In general, the corresponding spatial patterns are complex, but well-defined objects are often seen which evidently arise from individual localized states. These objects interact, and at times form a lattice-like arrangement.

Recently, various scanning probe experiments^{6,7,8} have been used to investigate the quantum Hall state on a micrometre scale, with some intriguing structure observed down to the 100-nm level⁷. We employ a single-electron transistor (SET) fabricated on the tip of a tapered glass fibre^{6,9} as a scanning electrometer. The resistance of the SET is sensitive to the charge induced on its central island. The SET is placed above the surface of a GaAs/AlGaAs heterostructure that has a two-dimensional electron gas (2DEG) buried 100 nm below the surface. Moving the SET along the surface changes its resistance in response to the local electrostatic potential arising from the heterostructure. We map this potential by recording the feedback voltage (V_{fb}) applied to the heterostructure that keeps the induced charge on the central island of the SET constant⁶, and thus also its resistance.

This heterostructure potential mainly comes from fixed charges (such as donors) above the 2DEG. We focus on the change in this potential ΔV that occurs when electron density n is changed (by voltage on the backgate). There are two expected contributions to ΔV . The main part is the local 2DEG contact potential which tracks the variation of the Fermi level with n . There may also be potential changes resulting from the charging of the localized states.

Under the quantum Hall conditions, as n is varied the ΔV changes rapidly in a step-like manner at n values that produce integer filling factors ν (defined as $\nu = nh/Be$ where B is the magnetic field, and h/e is the flux quantum), where the Fermi level passes between Landau levels. We have used⁶ these potential steps as a marker to determine the spatial dependence of ν and, hence, density within the 2DEG. Now, employing higher resolution in position and density, we look for individual localized states in regions of almost-integer filling factor.

In our experiments, B is set at a fixed value near $\nu = 2$ or $\nu = 1$. We place the SET at a fixed, arbitrary point and record V_{fb} , the d.c. measurement of ΔV ; the electron density is swept by a voltage V_{bg} applied to the backgate. We also use V_{bg} to modulate n at a frequency above the roll-off of the feedback loop and we record the equivalent ‘transparency’ signal, the a.c. potential seen by the SET, which has a better signal-to-noise ratio. Alternatively, instead of sweeping density, we record a set of spatial images of these signals at various

fixed V_{bg} . As will be seen, these signals show marked variation on spatial scales down to ~ 100 nm and, notably, for density changes of less than 1 electron μm^{-2} .

Broadly, the density dependence of the signals tends to be one of two types, as shown in the examples in Fig. 1a for $\nu = 2$. Both d.c. signals, taken at different locations, display the expected step of $\sim \hbar\omega_c/2\pi$, where ω_c is the cyclotron frequency, at $\nu = 2$ (refs 6 and 10). The amplitudes of the associated transparency signals, however, differ greatly. The smaller signal in the right panel quantitatively corresponds to the derivative of the d.c. step, unlike the much larger signal in the left panel. Such an excessive transparency signal indicates that an equilibrium charge density is not established (at 70 Hz), owing to resistive effects. This is supported by detection of an out-of-phase signal component. We now consider only cases where the transparency shows equilibrium charging behaviour.

Figure 1b shows a representative set of such transparency curves taken at a grid of points spaced by $0.1 \mu\text{m}$. No accompanying out-of-phase signal is seen. The middle of the associated d.c. steps corresponds to a local filling factor $\nu = 2$, giving a local density of $2B/\hbar$

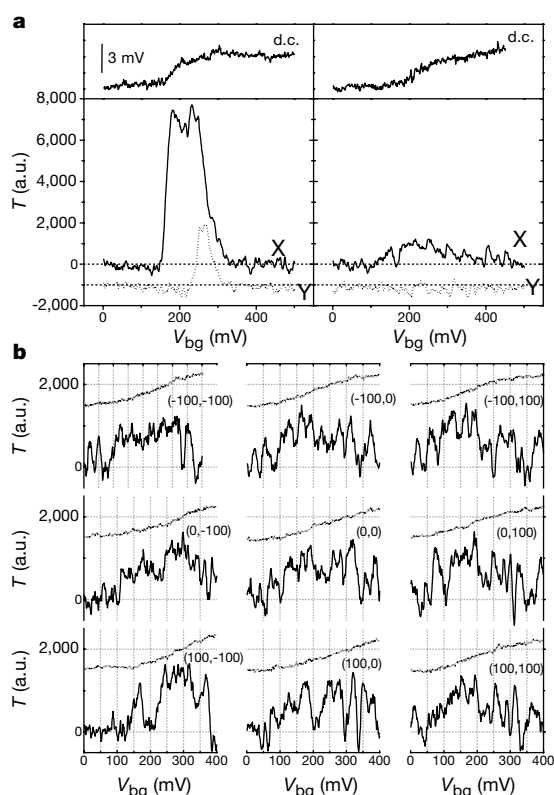


Figure 1 The d.c. and transparency signals at fixed points. **a**, Equilibrium versus non-equilibrium charging. Two typical sets of data (left and right panels) acquired by the single-electron transistor as the electron density ($\sim V_{bg}$) is swept through $\nu = 2$ at different spatial locations. The electron density n is $\sim 1.5 \times 10^{11} \text{ cm}^{-2}$, and varies by a few per cent with location. A backgate bias $V_{bg} = -1$ V reduces n by $1.3 \times 10^{10} \text{ cm}^{-2}$. Upper traces: d.c. signal displaying the characteristic step of the contact potential as the Fermi level moves between the Landau levels. Lower traces: in-phase component (X) and out-of-phase component (Y) (shifted down by 1,000) of transparency signal. The large transparency on the left panel is characteristic of non-equilibrium charging at 70 Hz. The smaller, completely in-phase transparency on the right panel signifies equilibrium charging. **b**, Rapid variation of transparency with density and position. The d.c. (upper traces) and in-phase transparency (lower traces) signals are plotted as a function of V_{bg} measured at nine spatial locations arranged in a 3×3 grid over an area of $0.2 \times 0.2 \mu\text{m}^2$. The middle of the d.c. steps serve as an estimate for the local $\nu = 2$ condition. Multiple oscillations seen in transparency signal are reproducible. No out-of-phase transparency signal is detected. All measurements shown are carried out at a temperature of 0.8 K.

($\hbar/e$) at the respective values of V_{bg} . All of these transparency signals display multiple oscillations with peaks separated by 20–40 mV in V_{bg} . Such patterns are largely reproducible in successive measurements, although they change slowly over time. In places, the transparency oscillations reach definite, reproducible, negative values. The spatial correlation length for the individual oscillation peaks is short and hard to quantify. Where peaks can be followed between adjacent points, they seem to shift in V_{bg} . As noted above, the integrated transparency signals reproduce the respective d.c. steps, but with a superior signal-to-noise ratio. In the integration, the oscillations result in small (~ 0.5 mV) step-like corrugations superimposed on the broader d.c. step of $\hbar\omega_c/2\pi$. The measured d.c. signals sometimes show this multiple-step structure, but usually it is obscured by noise. At a given point, these oscillations in the transparency are observed over a total density change of ~ 2 –4%, or 30–60 electrons μm^{-2} .

The spatial maps of the transparency in such regions usually show rather complex patterns that change rapidly with density. However, if the oscillations with density are well defined, there is often a simpler spatial pattern. An example is shown in Fig. 2. The transparency of the $0.5 \times 0.5 \mu\text{m}^2$ scanned area is shown for five successive additions of ~ 1.3 electrons μm^{-2} . There are clearly seen ring-like structures contracting into a common centre. The pattern almost repeats over three cycles of oscillation as the density is increased (only part of the evolution is shown), with a period of ~ 4 electrons μm^{-2} . Taken together, the ring size and the density period suggest that this cyclic behaviour involves the addition of a single electronic charge to the associated area. Clearly, such patterns provide the spatial visualization for the ‘sliding’ peaks seen in the density curves. We have seen many such contracting ring patterns, which more often show one to two cycles of repetition, but no clear examples of expanding rings. The rings and other associated patterns may appear anywhere across the region of the step in d.c. potential, both in the steeper part and in the tails.

The more complex transparency patterns often appear to contain multiple interacting and distorted ring-like objects. Particularly clear are the networks, such as appear in some of the $1 \times 1 \mu\text{m}^2$ panels of Fig. 3. These are taken near $\nu = 1$, at increments of 0.65 electrons μm^{-2} . The individual cells tend to be hexagonal, but not identical. In the sequence, the network is initially well formed, then falls apart. At a higher density, a new, denser pattern of cells forms in the same area (Fig. 3e), which again persists over a short range of density. The density of cells is ~ 15 –30 μm^{-2} , consistent with the charge density given by the estimated departure from integer filling

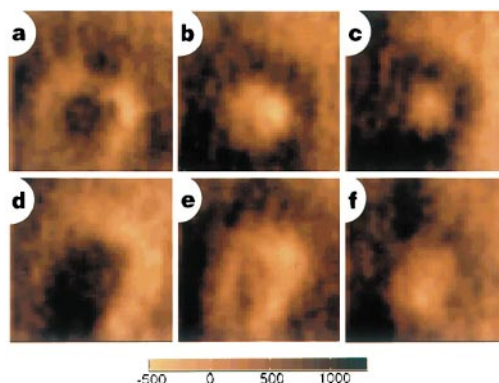


Figure 2 Spatial images of the transparency at $\nu = 2$ showing single-electron rings. **a–f**, Images of $0.5 \times 0.5 \mu\text{m}^2$ area are taken at six increasing densities. V_{bg} is changed by 10 mV between images which adds ~ 0.3 electron charges to the area of the panel. The transparency signals on the rings (light areas) are comparatively small, and sometimes negative (an overscreening of the a.c. backgate voltage), whereas those in the centre of the rings (dark regions) are larger and positive (an underscreening of the signal).

factor. This further suggests a direct relation between the cells and single-electron states.

Figure 4 shows the evolution of a portion of such a network in which the transition between the cells and isolated rings is seen. The successive images involve an addition of ~ 0.16 electrons to the area shown. Initially the structure has nearly triangular cells, but they change to a more hexagonal shape in Fig. 4b. In Fig. 4c and d the pattern shifts and the individual cells change in size. The cells separate as they shrink further in size (Fig. 4e), and they take on a ring-like form, a process which continues in Fig. 4f. The visual impression is that the rings are structures that have an individual identity. They have an elastic nature, in that they flatten out by mutual short-range repulsion when brought into contact, and take a circular shape when not actually touching. They have not been seen to merge. Many similar network patterns have been seen, both near $\nu = 1$ and, in less detail, near $\nu = 2$.

We took special care to test whether these spatial patterns could be induced by a gating effect of the tip. In all the experiments shown, the d.c. bias on the SET is set to nullify the contact potential. To do this, we adjust the voltage between the SET and the 2DEG to minimize the charge induced on the SET as the tip is lowered to the working height. During the spatial imaging, the effective gating voltage of the SET is less than 30 mV, ensuring that the amount of charge induced on the 2DEG by the SET is less than 0.5 electrons. In the test, the imaging of an area containing rings is then repeated with the tip biased off the original setting by 100–200 mV, so that the SET induces additional electrons or holes in the 2DEG. Such biased images display a different phase of the ring transformation but, qualitatively, the evolution of the ring pattern with density remains intact. We conclude that the ring patterns are not noticeably disturbed by the scanning tip. In addition, it seems improbable that well-developed networks of interacting rings such as occur in Figs 3 and 4 could be produced by a tip-gating effect.

The multiple oscillations of the transparency observed at a fixed spatial point are reminiscent of single-electron effects in quantum dots. There the analogous Coulomb-blockade charging behaviour is step-like changes in charge and an accompanying saw-tooth pattern of d.c. potential. When differentiated with respect to density, the d.c. saw-tooth yields an oscillating a.c. signal with negative regions, as indeed is seen here. Other experimental facts also strongly suggest

that single-electron charging is involved. First, as mentioned above, the period of the cyclic contraction of the rings approximately corresponds to the addition of a single-electron charge over the area of the structure. Second, charge variations of $\sim e$ over a typical area of the rings are required to produce the multiple d.c. steps of the observed ~ 0.5 mV magnitude.

However, the complete set of properties cannot be accounted for by assuming the discrete charging of isolated localized states. We see no well-defined 'quantum dots' of constant shape. Although in some manner the collapsing of a ring through one cycle adds one electron to the immediate area, how this takes place is not clear. A Coulomb blockade model in which the electron is transferred into a fixed potential minimum created by disorder does not provide a continuous spatial variation in signal. A possible way to improve the fixed potential model is by incorporating interaction with other available electrons. It is broadly accepted, however, that the density of electrons available for screening is determined by the deviation from integer filling factor^{11,12}. The total density range of the non-zero transparency is about $\sim 3 \times 10^9 \text{ cm}^{-2}$, yielding an average distance between screening electrons of more than 200 nm. This distance is comparable to typical sizes of the ring patterns and to the distance between the cells in Figs 3 and 4. It is not clear how the continuous variation of the potential required by the ring can be provided with so few 'free' carriers. It is also difficult to understand how the network of interacting rings can develop within this model. For example, the positions of the cells in the network pattern seem to be dictated entirely by the neighbouring cells.

The continuous spatial variation of the rings, and even more so the network patterns, suggests an analogy with the continuous modulation of density that occurs in the Wigner crystal or in a charge-density wave. In such a case, the disorder potential is of secondary importance, primarily causing pinning, whereas the electron interactions define the overall evolution of the patterns. Indeed, we do not observe any correlation between the ring centres and the extrema of the surface potential. So far we are not able to develop this analogy to a reasonable model. Although it is tempting to identify the network with a small Wigner crystallite, this is premature. The ordering of the Wigner crystal originates from long-range interaction between electrons. The interaction between the cells of our networks appears to be of short range, implying that the cells (and the rings) are dipole objects. □

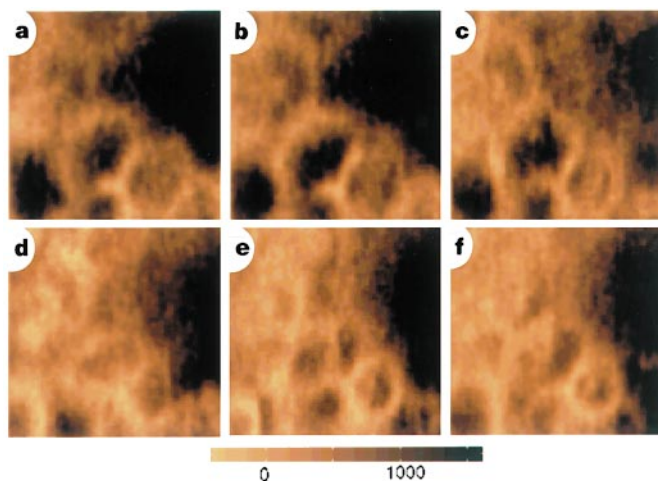


Figure 3 Interacting rings. **a–f**, Spatial images of the transparency signal over $1.0 \times 1.0 \mu\text{m}^2$ area taken at six increasing densities. V_{bg} is changed by 5 mV between images which adds ~ 0.65 electron charges to the area of the panel. Filling factor is close to 1. As in Fig. 2, the transparency signals are large and positive in the centre of the cells, and small, or even negative, on the cell boundaries.

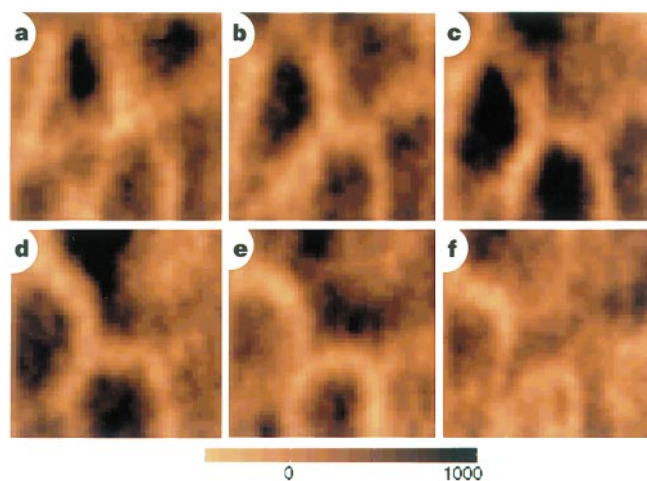


Figure 4 Close-up of colliding rings. **a–f**, Spatial images of the transparency signal over an area of $0.5 \times 0.5 \mu\text{m}^2$ taken at six increasing densities. V_{bg} is changed by 5 mV between images which adds ~ 0.16 electron charges to the area of the panel. Filling factor is close to 1.

Received 22 June 1999; accepted 6 January 2000

1. Prange, R. E. & Girvin, S. M. (eds) *The Quantum Hall Effect* (Springer, New York, 1990).
2. Laughlin, R. B. Quantized Hall conductivity in two dimensions. *Phys. Rev. B* **23**, 5632–5634 (1981).
3. Shklovskii, B. I. & Efros, A. L. *Electronic Properties Of Doped Semiconductors* (Springer, Berlin, 1985).
4. Kastner, M. A. Artificial Atoms. *Phys. Today* **46**, 24–31 (1993).
5. Yoo, M. J. *et al.* Scanning single-electron transistor microscopy: imaging individual charges. *Science* **276**, 579–582 (1997).
6. Yacoby, A. *et al.* Electrical imaging of the quantum Hall state. *Solid State Commun.* **111**, 1–13 (1999).
7. Tessler, S. H. *et al.* Subsurface charge accumulation imaging of the quantum Hall liquid. *Nature* **392**, 51–55 (1998).
8. McCormick, K. L. *et al.* Scanned potential microscopy of edge and bulk currents in the quantum Hall regime. *Phys. Rev. B* **59**, 4654–4657 (1999).
9. Fulton, T. A. & Dolan, G. J. Observation of single-electron charging effects in small tunnel junctions. *Phys. Rev. Lett.* **59**, 109 (1987).
10. Smith, T. P. *et al.* Direct measurement of the density of states of a two-dimensional electron gas. *Phys. Rev. B* **32**, 2696–2699 (1985).
11. Efros, A. L. Density of states of 2D electron gas and width of the plateau of IQHE. *Solid State Commun.* **65**, 1281 (1988).
12. Efros, A. L. Non-linear screening and the background density of states of 2DEG in magnetic field. *Solid State Commun.* **67**, 1019–1022 (1988).

Acknowledgements

We thank L. Levitov, S. Simon, L. Balents and C. Varma for useful discussions.

Correspondence and requests for materials should be addressed to N. Z. (e-mail: zhiten@physics.bell-labs.com).

Anisotropic spinodal dewetting as a route to self-assembly of patterned surfaces

A. M. Higgins & R. A. L. Jones

Department of Physics and Astronomy, University of Sheffield, The Hicks Building, Sheffield S3 7RH, UK

The ability to pattern surfaces on a microscopic length scale is of importance for technological applications such as the fabrication of microelectronic circuits and digital storage media. Devices fabricated entirely from polymers are now available, opening up the possibility of adapting polymer processing technologies to fabricate cheap, large-area devices using non-lithographic techniques^{1,2}—for example, by exploiting dewetting³ and phase separation^{4–6} in thin films. But the final pattern adopted by the polymer film using such approaches requires a template printed onto the substrate by optical lithography, microcontact printing^{4,5} or vapour deposition³. Here we describe a simple process for patterning surfaces that does not require a template. Our method involves the spinodal dewetting of a polymer surface by a thin polymer film, in which a liquid film breaks up owing to the amplification of thermal fluctuations in film thickness induced by dispersion forces^{7–14}. A preferred orientation is imposed on the dewetting process simply by rubbing the substrate, and this gives rise to patterns of remarkably well-aligned polymer lines. The width of these lines is well-defined, and is controlled by the magnitude of the dispersion forces at the interface, which in turn can be varied by varying the thickness of the polymer substrate. We expect that further work will make it possible to optimize the degree of order in the final morphology.

Previous work has shown that thin films of poly(methyl methacrylate) (PMMA) on polystyrene (PS) substrates are unstable with respect to spinodal dewetting⁹. We prepared bilayers of PMMA on PS as follows. First, PS ($M_n = 248,900$, $M_w/M_n = 1.04$, Polymer Source Inc., Canada) was applied by spin-coating toluene solutions of this compound onto silicon substrates. (Here M_n and M_w are

respectively the number- and the weight-average molecular masses.) After removing residual solvent, the film thickness was measured by ellipsometry. Second, the PMMA layer was added by spin-coating a PMMA solution onto cleaved mica, and then floating the resulting film onto water and depositing this on top of the silicon/PS samples. Dewetting was observed by annealing the samples at 150 °C on a hot-stage in an optical microscope in reflection mode (Nikon ME600D). The experiments on isotropic samples were performed twice, using PMMA of two slightly different molecular masses ($M_n = 102,700$, $M_w/M_n = 1.09$, Polymer Source Inc.; $M_n = 138,256$, $M_w/M_n = 1.04$, Polymer Laboratories Ltd, UK). In each case, three different thicknesses of PS were used, a single PMMA film being deposited onto all three and also onto a bare silicon fragment with a known oxide thickness (this was used to measure the PMMA film thickness). Each sample was annealed until dewetted morphologies were clearly visible. In each case, annealing was stopped as soon as sufficient contrast was obtained to provide good quality images but before any coarsening of the morphology took place. For the samples annealed here this meant annealing times of one hour or less (significant coarsening does not occur until after several hours). For each thickness of PS, a silicon/PS sample was also annealed to check that this did not dewet on its own. On the timescales of the experiments, no dewetting was seen on any of the silicon/PS samples.

Figure 1a shows the typical morphology for a PS thickness of 1,750 Å. This is a characteristic pattern resulting from isotropic spinodal dewetting; the pattern is clearly characterised by a single length-scale, but is isotropic. This is made clearer by taking a Fourier transform of the image. To obtain fast Fourier transforms (FFTs) of the dewetted structures, photographs were taken at lower magnification ($\times 20$ objective) than that used in Fig. 1a, at a number of different places on each sample. Each photo was decomposed into red, green and blue colour channels, and FFTs were taken of each colour channel. The FFTs were then radially averaged. A typical FFT for the 1,750-Å PS sample is shown in Fig. 1b. The well-defined ring

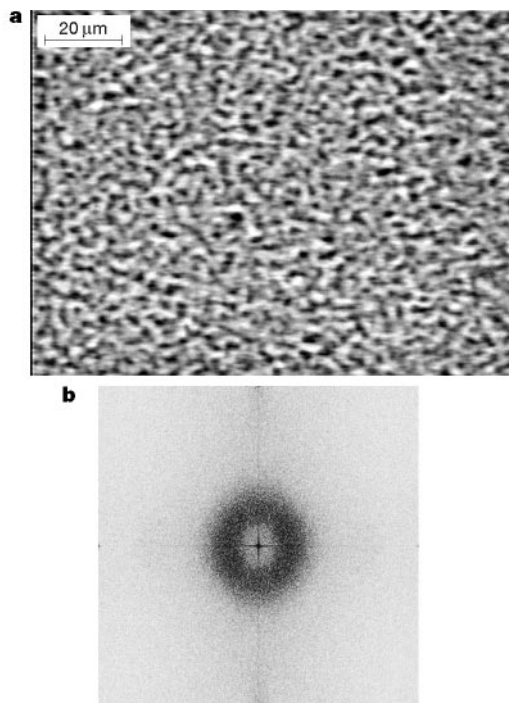


Figure 1 Dewetted morphology of a Si/PS/PMMA sample, where the PMMA film was prepared on cleaved mica. **a**, Optical micrograph of an annealed Si/PS/PMMA sample with a PS film thickness of 1,750 Å and a PMMA film thickness of 123 Å. **b**, Fast Fourier transform of a red-colour-channel image from this sample.

indicates the existence of a characteristic length-scale, while the radial symmetry of the transform confirms the isotropy of the dewetted pattern. The value of the characteristic wavelength was obtained for each sample by finding the maximum value of the radially averaged Fourier transform. These values are shown in Fig. 2 as a function of the thickness of the PS layer.

In order to obtain an oriented dewetted pattern, we prepared samples in exactly the same way as described above, except that the PMMA films were spin-coated onto glass slides (made from low-iron glass; BDH) that had previously been rubbed with lens tissue several times in one particular direction. Tapping-mode atomic force microscopy (AFM) studies (Digital Instruments MultiMode) of the substrate revealed that the surface topography of the glass is changed on the ångström length-scale by the rubbing process. Initially the surface is patterned with etch-pit-like structures about 30 nm across and 4 Å deep, with a typical separation of the order of 200 nm. Following rubbing, the etch pits are no longer apparent, and instead one sees patterns of raised ridges with an r.m.s. roughness of ~ 7 Å oriented in the rubbing direction. When PMMA is spin-coated onto rubbed substrates, the surface in contact with the substrate is imprinted with this topography. AFM pictures reveal a pattern of grooves with a depth of the order of 4 Å (Fig. 3a). A FFT of this pattern (Fig. 3b) shows a distinct degree of orientation, but rather than a single well-defined length-scale there is a broad distribution of lengths in the range 0.2–1 μm .

When samples prepared using the rubbed substrate were annealed following the same procedure as was used for isotropic systems, we found patterns exemplified by that shown in Fig. 4a, where the direction of rubbing was left to right across the image. The Fourier transforms, of which an example is shown in Fig. 4b, confirm the visual impression from the photomicrographs that the dewetting is now strongly anisotropic; instead of the random distribution of ramified droplets shown in Fig. 1a we have a series of more or less continuous lines of polymer running across the sample. Once again, the value of the characteristic length can be found from the Fourier transforms; these results are plotted in Fig. 2 together with the results for the isotropic samples. We note that the length-scales obtained for the oriented samples are identical (within error) to those obtained for the isotropic samples.

For both oriented and isotropic patterns, the wavelength of the dewetted morphology decreases as the PS layer becomes thinner. The linear theory of spinodal dewetting⁷ predicts that the fastest-growing unstable capillary wave has the wavevector $q_m = (|A|/4\pi h^4 \sigma)^{1/2}$, where σ is the interfacial tension, h is the film thickness and A is the Hamaker constant describing the dispersion forces acting across the thin PMMA layer between air and the substrate. In this case, the substrate is a composite of silicon coated with polystyrene. Both the relative permittivity and the refractive

index in the visible range for silicon are larger than those of either PS or PMMA. We therefore expect, on the basis of the Lifshitz theory of dispersion forces¹⁵, that the magnitude of the dispersion force between the air and the 'substrate' (silicon/PS combined) across the PMMA should increase as the PS thickness is reduced. Because it is clear from the experiments that the dispersion force between the air and the 'substrate' must be attractive, that is, with a negative value—otherwise dewetting would not occur—increasing the magnitude of this attractive force should give rise to a reduction in the size-scale of the dewetted morphology.

We believe that the strong orientation in the dewetted pattern arises from the modulation in PMMA thickness that comes from the process of rubbing the substrate. This modulation has a definite orientation and a broad range of wavelengths in the submicrometre range, which would lead to a weak spatial modulation in the dispersion force field which causes the interfacial instability. Spinodal dewetting provides a mechanism by which the effects of a weak perturbing field can be selectively amplified; those thickness fluctuations with the correct wavelength are picked out and magnified by the dewetting process. The result is that the orientation of the perturbing field is preserved in the final dewetted pattern, but the length-scale of the dewetting remains identical to the isotropic case. This mechanism has some analogies to the way that, in spinodal decomposition, the presence of a surface can lead to phase separation, resulting in an ordered lamellar structure. In such surface-directed spinodal decomposition, a perturbing field imposes a well-defined one-dimensional ordering on what otherwise would be an isotropic three-dimensional pattern, but once again the

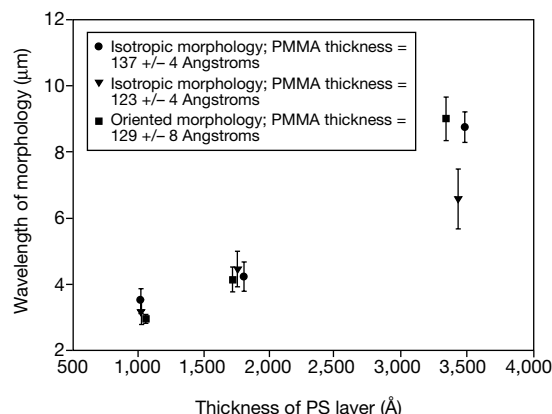


Figure 2 The length-scale of the dewetted morphology as a function of PS thickness for both isotropic and oriented morphologies.

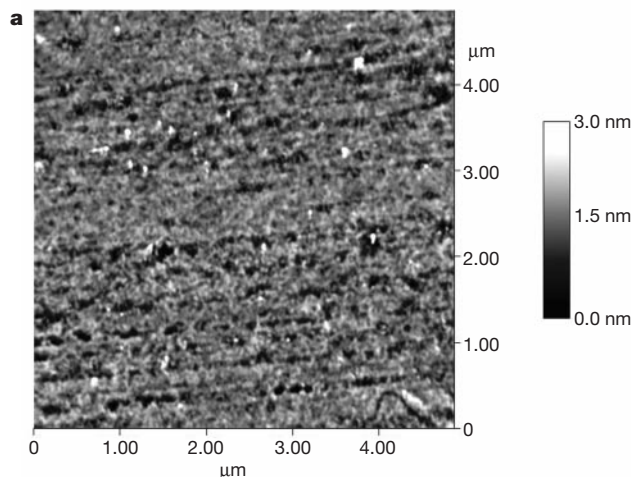


Figure 3 Morphology of underside of PMMA film floated from rubbed glass. **a**, Topography obtained using AFM. **b**, Fast Fourier transform of **a**.

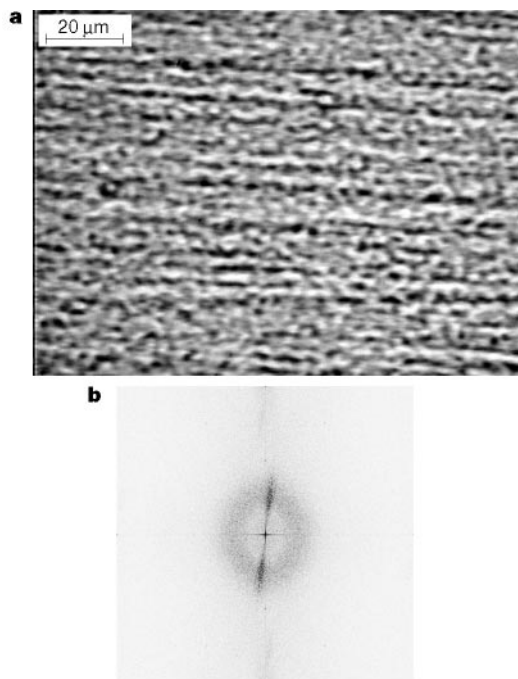


Figure 4 Dewetted morphology of Si/PS/PMMA sample, where the PMMA film was prepared on rubbed glass. **a**, Optical micrograph of an annealed Si/PS/PMMA sample with a PS film thickness of 1,710 Å and a PMMA film thickness of 129 Å. **b**, Fast Fourier transform of a red-colour-channel image from this sample.

presence of a wavelength selection mechanism ensures that the length-scale of the morphology is the same in both the bulk isotropic situation and the surface-directed lamellar case¹⁶.

The process of spinodal dewetting belongs to a general class of phenomena, including spinodal decomposition and viscous fingering, in which noise—whether equilibrium thermal noise or random variations originating from the history of the sample—is amplified. Because the amplification mechanism is wavelength selective (even though the initial noise is present at a very wide range of length-scales), the final pattern, while random, is characterized by a single length. Our experiments show that a high degree of order may be produced in the final morphology by imposing a rather subtle directional bias on the pre-existing fluctuations. The degree of perfection of the final pattern presumably depends on the strength of this perturbing bias relative to the pre-existing fluctuations; theory, simulation and further experimental work on this technique should make it possible to achieve oriented morphologies with a greater degree of perfection. □

Received 27 August 1999; accepted 12 January 2000.

1. Friend, R. H. *et al.* Electroluminescence in conjugated polymers. *Nature* **397**, 121–128 (1999).
2. Service, R. F. Patterning electronics on the cheap. *Science* **278**, 383–384 (1997).
3. Gau, H., Herminghaus, S., Lenz, P. & Lipowsky, R. Liquid morphologies on structured surfaces: From microchannels to microchips. *Science* **283**, 46–49 (1999).
4. Böltau, M., Walheim, S., Mlynek, J., Krausch, G. & Steiner, U. Surface-induced structure formation of polymer blends on patterned substrates. *Nature* **391**, 877–879 (1998).
5. Heier, J., Kramer, E. J., Walheim, S. & Krausch, G. Thin diblock copolymer films on chemically heterogeneous surfaces. *Macromolecules* **30**, 6610–6614 (1997).
6. Nisato, G., Ermi, B. D., Douglas, J. F. & Karim, A. Excitation of surface deformation modes of a phase-separating polymer blend on a patterned substrate. *Macromolecules* **32**, 2356–2364 (1999).
7. Brochard-Wyart, F. & Daillant, J. Drying of solids wetted by thin liquid films. *Can. J. Phys.* **68**, 1084–1088 (1990).
8. Xie, R., Karim, A., Douglas, J. F., Han, C. C. & Weiss, R. A. Spinodal dewetting of thin polymer films. *Phys. Rev. Lett.* **81**, 1251–1254 (1998).
9. Sferrazza, M. *et al.* Interfacial instability driven by dispersive forces: the early stages of spinodal dewetting of a thin polymer film on a polymer substrate. *Phys. Rev. Lett.* **81**, 5173–5176 (1998).
10. Reiter, G. Dewetting of thin polymer films. *Phys. Rev. Lett.* **68**, 75–78 (1992).
11. Krausch, G. Dewetting at the interface between two immiscible polymers. *J. Phys. Condens. Matter* **9**, 7741–7752 (1997).

12. Lambooy, P., Phelan, K. C., Haugg, O. & Krausch, G. Dewetting at the liquid-liquid interface. *Phys. Rev. Lett.* **76**, 1110–1113 (1996).
13. Stange, T. G., Evans, D. F. & Hendrickson, W. A. Nucleation and growth of defects leading to dewetting of thin polymer films. *Langmuir* **13**, 4459–4465 (1997).
14. Herminghaus, S. *et al.* Spinodal dewetting in liquid crystal and liquid metal films. *Science* **282**, 916–919 (1998).
15. Israelachvili, J. *Intermolecular and Surface Forces* (Academic, London, 1992).
16. Jones, R. A. L., Norton, L. J., Kramer, E. J., Bates, F. S. & Wiltzius, P. Surface-directed spinodal decomposition. *Phys. Rev. Lett.* **66**, 1326–1329 (1991).

Acknowledgements

We thank K. Dalnoki-Veress and M. Sferrazza for help and advice with sample preparation and data analysis. We also thank G. Reiter for the suggestion that the magnitude of the Hamaker constant could be modified by varying the thickness of the lower polystyrene film. This work was supported by the EPSRC through the ROPA programme.

Correspondence and requests for materials should be addressed to R.A.L.J. (e-mail: R.A.L.Jones@Sheffield.ac.uk).

A soluble and air-stable organic semiconductor with high electron mobility

H. E. Katz, A. J. Lovinger, J. Johnson, C. Kloc, T. Siegrist, W. Li, Y.-Y. Lin & A. Dodabalapur

Bell Laboratories-Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA

Electronic devices based on organic semiconductors offer an attractive alternative to conventional inorganic devices due to potentially lower costs, simpler packaging and compatibility with flexible substrates^{1,2}. As is the case for silicon-based microelectronics, the use of complementary logic elements—requiring n- and p-type semiconductors whose majority charge carriers are electrons and holes, respectively—is expected to be crucial to achieving low-power, high-speed performance. Similarly, the electron-segregating domains of photovoltaic assemblies require both n- and p-type semiconductors^{3–5}. Stable organic p-type semiconductors are known⁶, but practically useful n-type semiconductor materials have proved difficult to develop, reflecting the unfavourable electrochemical properties of known, electron-demanding polymers⁷. Although high electron mobilities have been obtained for organic materials, these values are usually

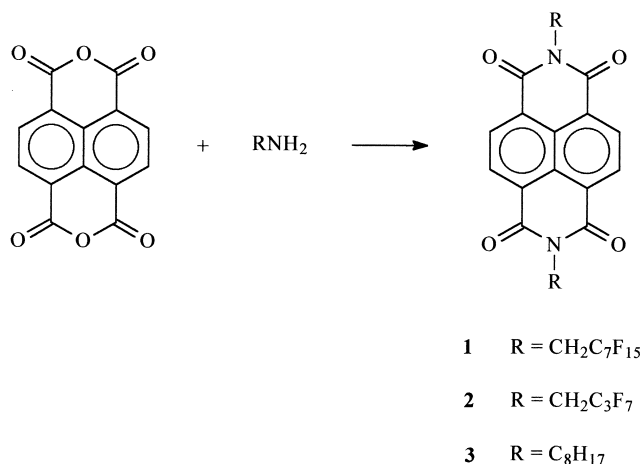


Figure 1 Synthesis and structures of naphthalenetetracarboxylic diimide compounds 1–3.

obtained for single crystals at low temperatures, whereas practically useful field-effect transistors (FETs) will have to be made of polycrystalline films that remain functional at room temperature. A few organic n-type semiconductors that can be used in FETs are known, but these suffer from low electron mobility, poor stability in air and/or demanding processing conditions^{8–10}. Here we report a crystallographically engineered naphthalenetetracarboxylic diimide derivative that allows us to fabricate solution-cast n-channel FETs with promising performance at ambient conditions. By integrating our n-channel FETs with solution-deposited p-channel FETs, we are able to produce a complementary inverter circuit whose active layers are deposited entirely from the liquid phase. We expect that other complementary circuit designs¹¹ can be realized by this approach as well.

Compound **1**, and its shorter and its non-fluorinated analogues, **2** and **3** respectively (Fig. 1), were synthesized by conventional means¹². Amines and 1,4,5,8-naphthalenetetracarboxylic dianhydride were heated in quinoline in the presence of zinc acetate catalyst, and the products crystallized, collected, washed with hot aqueous Na₂CO₃, water, and toluene, and sublimed under vacuum. Elemental analyses were within 0.26 of theoretical percentages. Thin films were initially vacuum-deposited onto SiO₂–Si substrates and carbon grids. The powder and thin-film crystallinity of **1** and **3** were excellent, while **2** was somewhat more disordered (Fig. 2). Films of all the compounds showed X-ray and electron diffraction patterns consistent with the molecular long axes being nearly perpendicular to the substrate plane. In this orientation, the π systems are arranged so that electron transfer between ring systems in directions parallel to the substrate would be favourable. Single crystals of **1** were grown by slowly cooling and concentrating saturated xylene solutions. Based on preliminary diffraction data, the molecules are in a dense, two-dimensional 'herringbone' arrangement (Fig. 3)

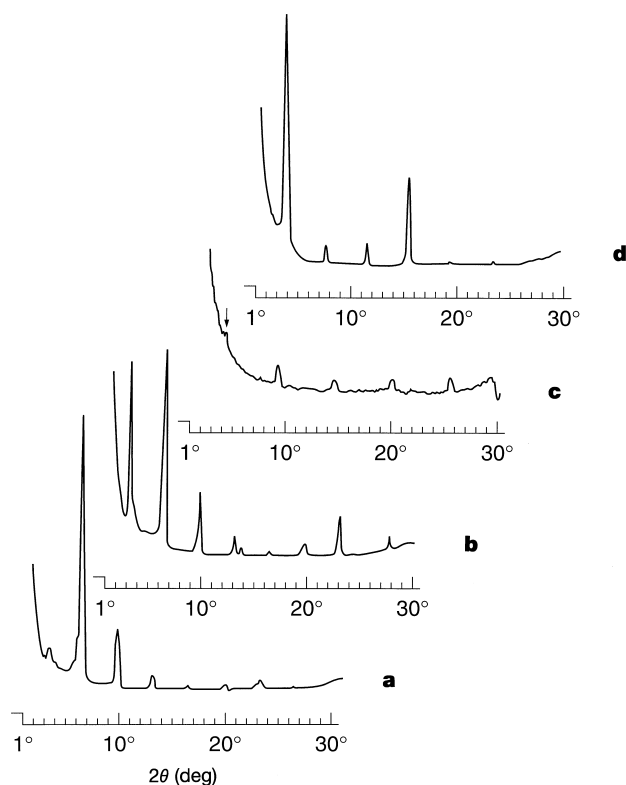


Figure 2 X-ray diffraction patterns of films deposited onto SiO₂. The substrates were held at the given temperatures, and first order peaks were observed corresponding to the given spacings: **a**, **1** sublimed at 70 °C, 26 Å; **b**, **1** cast from trifluorotoluene at 100 °C, 27 Å; **c**, **2** sublimed at 50 °C, 17 Å; **d**, **3** sublimed at 68 °C, 23 Å. Arrow indicates the first-order diffraction peak.

reminiscent of p-channel compounds such as α -6T and α -8T (sexithiophene and octithiophene)^{13–15}, and associated with considerable charge carrier mobility. The half-unit-cell length of 26.6 Å is within 1 Å of the first-order X-ray diffraction peak of films on SiO₂.

Mobilities were measured in the saturation regime¹⁶ on FETs where the semiconducting films were deposited at substrate temperatures of 30–100 °C, with 3,000-Å-thick SiO₂ dielectrics, and evaporated gold top-contact source and drain electrodes. Mobilities for **1** were up to 0.06 cm² V^{−1} s^{−1} for a width/length ratio (W/L) of 17, and 0.1 cm² V^{−1} s^{−1} for $W/L = 1.5$, for operation in air after deposition at 70 °C (Fig. 4). On/off ratios (100 V/0 V gate) above 100,000 were obtained. The devices could be cycled repeatedly with 0–50 V applied to drain and gate relative to source, until failure of the gold contacts (possibly because of delamination or chemical degradation caused by resistive heating at the gold–semiconductor junction). Active layers at least as thin as 150 Å could be used, and a photocurable polyimide was suitable as an alternative dielectric. Films deposited on Dow BCB dielectric, on the other hand, were inactive, possibly because of conjugated carbonyl traps formed on curing the polymer. Currents were two orders of magnitude lower when bottom-contact gold was used instead of top contacts, but were only a factor of two lower when bottom contacts fabricated from carbon ink were used. Although uniform films with the highest activity were obtained at deposition rates of $\sim 2 \text{ Å s}^{-1}$, a film with mobility of 0.003 cm² V^{−1} s^{−1} was produced by a relatively uncontrolled 5-s sublimation.

The mobility of **2** was $\sim 0.01 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ when deposited at 25–50 °C, and showed an on/off ratio of 40,000. The compound was volatile enough to desorb from substrates heated to higher temperatures. Compound **3**, which has been used as an electron acceptor in photorefractive composites¹⁷, showed no FET activity in air, even though its films were highly crystalline. However, degassing a sample (70 °C deposition temperature) in a turbo-pumped chamber and testing under high vacuum resulted in a progressively increasing mobility: 0.001 in 6 h, 0.01 in 1 day, 0.08 in

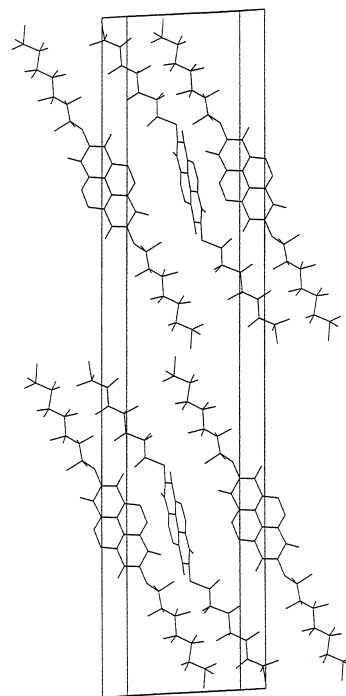


Figure 3 Crystal packing diagram of **1** showing the herringbone motif. The unit cell parameters (C-centred monoclinic) are $a = 53.22 \text{ Å}$, $b = 5.20 \text{ Å}$, $c = 11.81 \text{ Å}$, $\beta = 92.79^\circ$, and $Z = 4$.

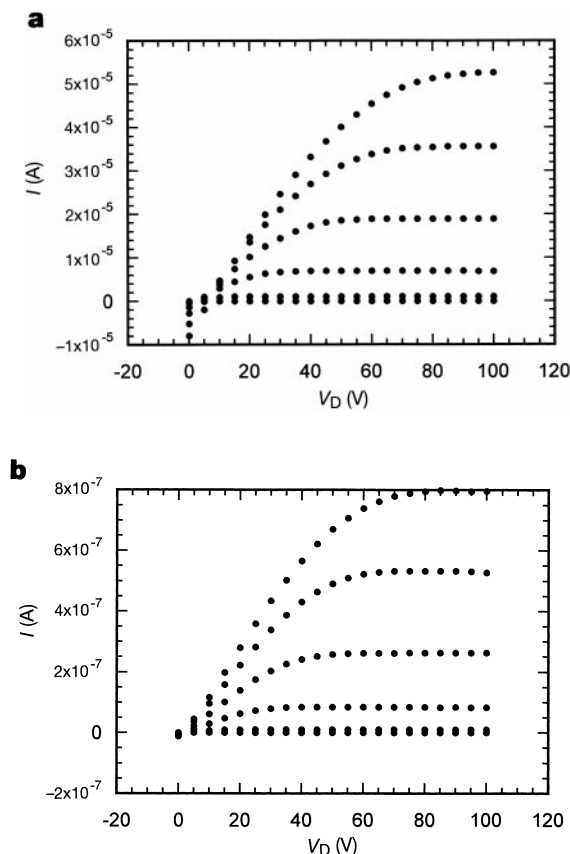


Figure 4 FET characteristics for films of **1** prepared by sublimation and solution deposition. **a**, The film was sublimed onto Si/SiO₂ held at 70 °C. The channel length was 0.25 mm, and the width was 4.5 mm. **b**, The film was cast from a 200-p.p.m. solution in trifluorotoluene onto Si/SiO₂ at 100 °C, length 0.15 mm, width 0.25 mm. Gate voltages were 0–100 V from bottom to top curves in increments of 20 V. V_D is the drain-source voltage.

3 days, and $0.16 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ in 4 days using contacts with a W/L of 17. This is, to our knowledge, the highest mobility yet observed for an unmodified organic material in an n-channel thin film FET. The admission of oxygen into the chamber lowered the mobility by 1–2 orders of magnitude within minutes, and introduction of air rapidly deactivated the device, presumably because of ambient humidity. The electronic effect of the fluorinated chains may be important in stabilizing **1** and **2** relative to **3** during operation in air; the electrochemical reduction potential of **1** in tetrahydrofuran is 150 mV less negative than that of **3**. However, the denser packing of the fluorinated chains compared to what would be expected in **3** is probably a more important factor. Based on space-filling models and Fig. 3, the fluorinated chains are separated by channels $\sim 2 \text{ \AA}$ on a side. Hydrocarbon chains similarly disposed would leave $4\text{-}\text{\AA}$ channels, which could well be more permeable to oxygen and water. Other, more subtle crystal packing differences may also be involved in stabilizing.

A variety of halogenated and aromatic solvents were screened as vehicles for liquid-phase deposition of films of **1**. Films were cast at elevated temperatures at concentrations of tens to hundreds of parts per million (p.p.m.). Solvent removal was occasionally vacuum-assisted. Patterns of rings or unconnected crystallites were often obtained, and such morphologies did not lead to active semiconductor films. When 200–400 p.p.m. solutions of **1** in α,α,α -trifluorotoluene were used at 100 °C, with solvent evaporation under ambient pressure, regions with FET mobility $>0.01 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ were obtained. Total active areas were $\sim 1 \text{ cm}^2$, although mobility was not uniform over the entire area. This result shows that liquid-phase deposition of n-channel semiconductors is feasible. Variations in

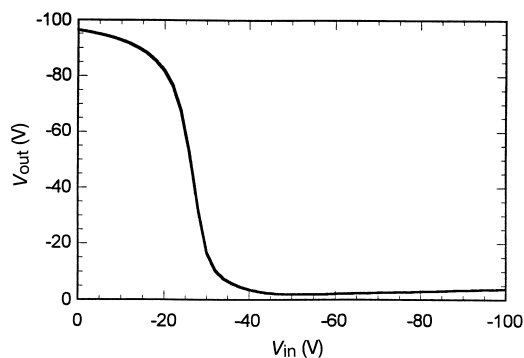


Figure 5 Inverter characteristics for solution-cast **1** and dihexyl- α -quinuethiophene. See text for details. The supply voltage is -100 V and the n-channel FET is the load.

the dielectric surface chemistry, the nature of the naphthalenetetracarboxylic diimide side chain, the means of applying the solution, and the temperature and pressure cycles during solvent evaporation, could potentially improve the film quality and the efficiency of film preparation.

We fabricated a p-channel FET from an α,ω -dihexylquinuethiophene-toluene solution^{6,18} (mobility $0.1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) on one portion of a substrate, and made a solution-cast FET with **1** on another part of the same substrate: these FETs shared the same gate electrode and each had separate gold source and drain top contacts. The input (common gate), output (connected to one top contact from each FET), supply, and load contacts (the remaining top contacts) were connected and biased in an inverter configuration¹⁹ (in which the output is expected to be the “opposite” of the input). Switching from $<5 \text{ V}$ to $>95 \text{ V}$ with a sensitivity of up to 10 was observed (Fig. 5), showing that this complementary inverter circuit worked successfully. □

Received 13 October 1999; accepted 15 February 2000.

- Matters, M. *et al.* Organic field-effect transistors and all-polymer integrated circuits. *Opt. Mater.* **12**, 189–197 (1999).
- Drury, C. J., Mutsaers, C. M. J., Hart, C. M., Matters, M. & de Leeuw, D. M. Low-cost all-polymer integrated circuits. *Appl. Phys. Lett.* **73**, 108–110 (1998).
- Yamaguchi, Y., Fujiyama, T. & Yokoyama, M. Bipolar-charge-transporting organic photoconductors. *J. Appl. Phys.* **70**, 855–859 (1991).
- Panayiotatos, P. Recent advances in optimally designed p-n organic semiconductor solar cells. *IEEE Photovoltaic Specialists Conf.* **19**, 889–894 (1987).
- Fox, M. A. Polymeric and supramolecular arrays for directional energy and electron transport over macroscopic distances. *Acc. Chem. Res.* **25**, 569–574 (1992).
- Li, W., Katz, H. E., Lovinger, A. J. & Laquindanum, J. G. Field-effect transistors based on thiophene hexamer analogues with diminished electron donor strength. *Chem. Mater.* **11**, 458–465 (1999).
- de Leeuw, D. M., Simenon, M. M. J., Brown, A. R. & Einerhand, R. E. F. Stability of n-type doped conducting polymers and consequences for polymeric microelectronic devices. *Synth. Met.* **87**, 53–58 (1997).
- Haddon, R. C. *et al.* C₆₀ thin film transistors. *Appl. Phys. Lett.* **67**, 121–123 (1995).
- Laquindanum, J. G., Katz, H. E., Dodabalapur, A. & Lovinger, A. J. n-Channel organic transistor materials based on naphthalene frameworks. *J. Am. Chem. Soc.* **118**, 11331–11332 (1996).
- Bao, Z., Lovinger, A. J. & Brown, J. New air-stable n-channel organic thin film transistors. *J. Am. Chem. Soc.* **120**, 207–208 (1998).
- Crone, B. *et al.* Large-scale complementary integrated circuits based on organic transistors. *Nature* **403**, 521–523 (2000).
- Rademacher, A., Maerle, S. & Langhals, H. Lösliche Perylen-Fluoreszenzfarbstoffe mit hoher Photostabilität. *Chem. Ber.* **115**, 2927–2934 (1982).
- Horowitz, G. *et al.* Growth and characterization of sexithiophene single crystals. *Chem. Mater.* **7**, 1337–1341 (1995).
- Siegrist, T. *et al.* The crystal structure of the high-temperature polymorph of α -hexathienyl (α -6T/HT). *J. Mater. Res.* **10**, 2170–2173 (1995).
- Fichou, D., Bachet, B., Demanze, F., Billy, I. & Horowitz, G. Growth and structural characterization of the quasi-2D single crystal of α -octithiophene (α -8T). *Adv. Mater.* **8**, 500–504 (1996).
- Horowitz, G. Organic field-effect transistors. *Adv. Mater.* **10**, 365–376 (1998).
- Wiederrecht, G. P., Miemczyk, M. P., Svec, W. A. & Wasielewski, M. R. Photorefractivity in nematic liquid crystals doped with a conjugated polymer: mechanisms for enhanced charge transport. *Chem. Mater.* **11**, 1409–1412 (1999).
- Katz, H. E. *et al.* Facile deposition processes for semiconducting molecular solids. *Proc. Mater. Res. Soc.* (in the press).
- Dodabalapur, A., Laquindanum, J., Katz, H. E. & Bao, Z. Complementary circuits with organic transistors. *Appl. Phys. Lett.* **69**, 4227–4229 (1996).

Acknowledgements

We thank Z. Bao, R. Raju, E. Chandross and J. Rogers for discussions. H.E.K. gratefully acknowledges discussions with the late Professor B. Weintraub.

Correspondence and requests for materials should be addressed to H.E.K. (e-mail: hek@lucent.com).

Molecular-scale interface engineering for polymer light-emitting diodes

Peter K. H. Ho^{††}, Ji-Seon Kim^{*}, Jeremy H. Burroughes[†], Heinrich Becker[‡], Sam F. Y. Li[§], Thomas M. Brown^{*}, Franco Cacialli^{*} & Richard H. Friend^{*†}

^{*} Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK

[†] Cambridge Display Technology, Greenwich House, Madingley Rise, Cambridge CB3 0HJ, UK

[‡] Covion Organic Semiconductors, Industrial Park Höchst G865, D-65926 Frankfurt am Main, Germany

[§] Department of Chemistry, National University of Singapore, and Institute of Materials Research and Engineering, Kent Ridge Crescent S119260, Singapore

Achieving balanced electron–hole injection and perfect recombination of the charge carriers is central to the design of efficient polymer light-emitting diodes^{1,2} (LEDs). A number of approaches have focused on modification of the injection contacts, for example by incorporating an additional conducting-polymer layer at the indium-tin oxide (ITO) anode^{3,4}. Recently, the layer-by-layer polyelectrolyte deposition route has been developed for the fabrication of ultrathin polymer layers^{5,6}. Using this route, we previously incorporated ultrathin (<100 Å) charge-injection interfacial layers in polymer LEDs⁷. Here we show how molecular-scale engineering of these interlayers to form stepped and graded electronic profiles can lead to remarkably efficient single-layer polymer LEDs. These devices exhibit nearly balanced injection, near-perfect recombination, and greatly reduced pre-turn-on leakage currents. A green-emitting LED comprising a poly(*p*-phenylene vinylene) derivative sandwiched between a calcium cathode and the modified ITO anode yields an external forward efficiency of 6.0 per cent (estimated internal efficiency, 15–20 per cent) at a luminance of 1,600 candelas per m² at 5 V.

Stable contacts with low interface resistance are required for high power throughput and long-term stability of semiconductor devices. This can be achieved in conventional III–V semiconductor technologies through composition grading, for example, to remove the interface energy offsets that hinder carrier motion⁸. Analogous concepts can be applied to polymer semiconductors if the desired polymers can be organized and then immobilized on the relevant carrier-hopping length-scale⁹ of 3–10 Å. The polyelectrolyte assembly process^{5,6} gives this capability. Polycations and polyanions are alternately adsorbed from their respective solutions onto the substrate to form a multilayer film in which the thickness and other physico-chemical properties of the sub-layers can be tuned over molecular widths^{5,10}. A strong cooperative ion-pair interaction locks the chain segments into highly interpenetrated structures¹¹. This affords mechanical robustness and solvent compatibility with subsequent processing steps.

Here we extend our previous work⁷ to fabricate functional stepped and graded electronic profiles in the charge-injection interlayer for a better control of hole injection and electron leakage at the contact between the ITO and the light-emitting polymer (LEP). To fabricate a stepped profile, we placed an additional 10-Å-

thin barrier of a partially conjugated poly(*p*-phenylene vinylene) (PPV) at the interface between the injection interlayer and the LEP. The injection interlayer comprises an alternating assembly of the anionic poly(3,4-ethylenedioxythiophene):poly(4-styrenesulphonate) (PEDT:PSS) complex with the cationic poly(*p*-xylylene- α -tetrahydrothiophenium) (PXT). To fabricate the graded profile, we partially de-doped the PEDT:PSS with hydrazine hydrate and then utilized this complex at various doping levels to construct the interlayer. This is illustrated schematically in Fig. 1. The PEDT:PSS complex is PSS-rich, leading to an overall negative charge density that is only fractionally increased on de-doping.

The electronic spectra shown in Fig. 2a confirm that the as-received saturation-doped PEDT:PSS (charge-per-ring ratio $\gamma = 0.3$) solutions can be successfully reduced¹² to a range of doping levels. (Here we will denote the various doping levels by p^{++} (saturation-doped), p^+ ($\gamma = 0.25$), p ($\gamma = 0.2$) and i ($\gamma = 0.02$).)

A schematic profile of the density of electronic states through the graded interlayer is shown in Fig. 2b. It illustrates how hole-transporting states can be arranged to bridge the Fermi level (E_F) of the ITO and the π -valence band edge of the LEP. The hole-injection barrier (Δ_h) for an abrupt ITO/LEP contact is related to the difference between the work function of the ITO and the ionization potential (I_p) of the LEP. When a uniformly p^{++} -doped PEDT interlayer is inserted between the ITO and the LEP, the Δ_h becomes related to the I_p difference between p^{++} -PEDT and the LEP. With progressive de-doping of this interlayer, its I_p can be increased by a few tenths of an eV (refs 13, 14) to permit a better match with that of the LEP. The gradation of I_p thus partitions an abrupt Δ_h barrier into a series of smaller steps, to allow an easier hopping injection of holes from ITO into the LEP. This is expected to be particularly useful for blue-emitting polymers with large I_p values (such as polyfluorene) for which hole injection is difficult in conventional device architectures. A second feature depicted in

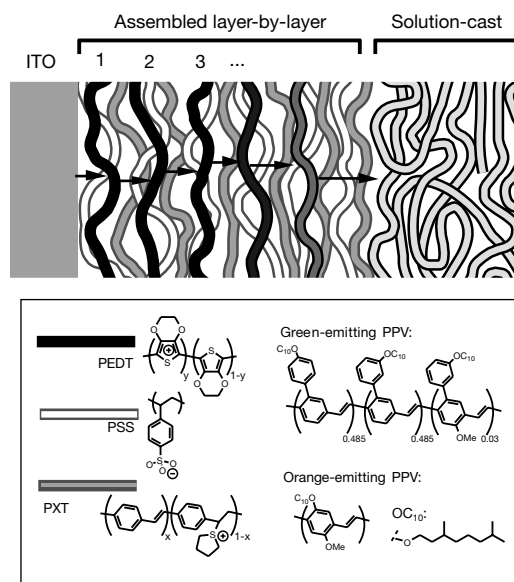


Figure 1 Schematic diagram of the hole-injection interlayer between ITO and the LEP, and chemical structure of some of the polymers. The interlayer depicted is a graded (PEDT:PSS/PXT)₅(PSS/PXT) film assembled layer-by-layer on silylated ITO via alternate polyanion–polycation deposition. The LEP film is formed by spin-casting. The upper panel indicates how the sub-layers (10–15 Å thin) of the interlayer are stratified and yet highly interpenetrated to facilitate carrier hopping between neighbouring PEDT layers in the direction of the arrows. The reduced shading of the PEDT chains to the right indicates a decreasing doping level across the graded interlayer. Thermal baking at 70–80 °C eliminates tetrahydrothiophene to form sulphonate crosslinks between PXT and PSS. The PXT contains short conjugated oligophenylenevinylene sequences.

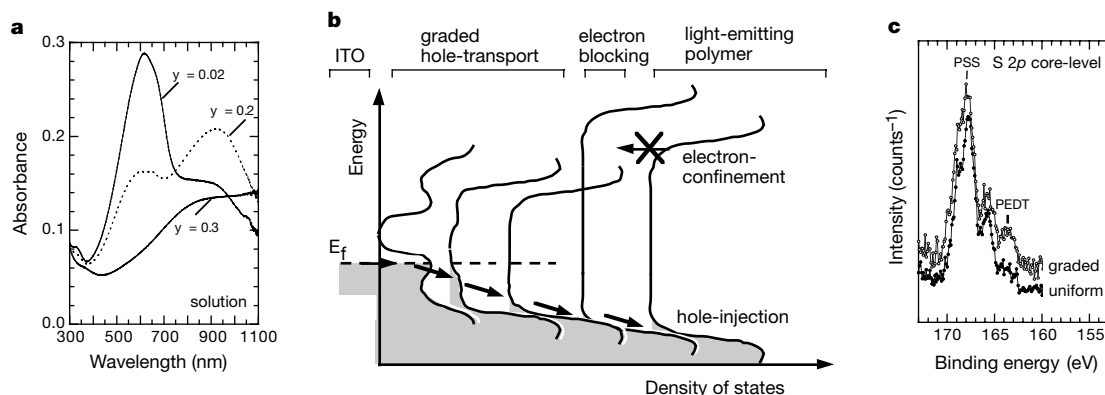


Figure 2 Schematic electronic profile of the graded interlayer and some experimental confirmation. **a**, Ultraviolet-visible-near infrared absorption spectra of PEDT:PSS in H₂O (same concentration) at different levels of dedoping by hydrazine. The intensity of the π – π^* band at 620 nm increases, while the polaronic band at >800 nm decreases, from which the doping level can be estimated. No precipitation was found. **b**, Schematic electronic density of states across a graded interlayer (occupied states are shaded). Holes are first injected into the gap states of the p^{++} -PEDT, then transported down less-doped PEDT states with larger I_p values and then injected into the LEP over a lower effective barrier. Also shown is the presence of an electron-blocking sub-layer fabricated from a

lower-electron-affinity polymer. This confines the injected electrons to help balance the electron and hole currents and give higher recombination efficiency with holes. **c**, S 2p core-level X-ray photoelectron spectra of (PEDT:PSS/PXT)₅(PSS/PXT) films on ITO assembled with uniformly-doped PEDT (5 p^{++} cycles) or graded-PEDT (2 p^{++} - p -i cycles). The binding energy is referenced to E_f . The presence of PSS (S 2p $_{3/2}$, 167.7 eV) and PEDT (S 2p $_{3/2}$, 163.5 eV) is confirmed. Spectral shape analysis of the PEDT component is consistent with the presence of undoped PEDT at the surface of the graded film.

Fig. 2b is the blocking of injected electrons by a sub-layer (for example, of partially conjugated PPV) that has a lower electron affinity than the LEP. This is thought to confine electrons to the LEP layer to help balance the electron and hole currents and give a better recombination efficiency with the injected holes.

The overall scheme therefore aims to integrate through molecular-scale assembly several elements previously known to improve LED performance (that is, hole-transport multilayer¹⁵, conducting polymer layer^{3,4}, and electron barrier⁷). To show that the target structures have been achieved, we will first outline some characterization results.

The S 2p core-level X-ray photoelectron spectra of a uniform (5 p^{++}) and a graded (2 p^{++} - p -i) (PEDT:PSS/PXT)₅(PSS/PXT) film on ITO are shown in Fig. 2c. (The numeral denotes the number of sub-layers assembled for each type.) The incorporation of PEDT and PSS is confirmed through their respective binding-energy peaks. In addition, there is a broadening of the PSS core-level (by 0.3 ± 0.1 eV) towards higher binding energies that is attributed to band bending within the sampling depth (~ 20 Å) of the graded film. Small mobile ions that could scramble the doping profile, such as Cl[–], were not incorporated into the film (< 0.1 atom%).

The intensity ratio of the C 1s to In 3d photoelectrons for the uniform film indicates a thickness of ~ 60 Å. Ellipsometry also gives ~ 60 Å ($\Delta(\Delta) = -18^\circ \pm 1^\circ$) for this film, and ~ 50 Å ($\Delta(\Delta) = -15^\circ \pm 1^\circ$) for the graded film. $\Delta(\Delta)$ is the change in the difference of the reflection phase shift between the p-polarized light and the s-polarized light measured at 275-nm wavelength. The excellent agreement confirms that well-defined polyelectrolyte monolayers, rather than large colloidal particles, have been deposited. This is further supported by the root-mean-square roughness values (1×1 μ m, by atomic force microscopy) of 40–45 Å for these films, not much larger than that of bare ITO (35–40 Å).

The slight decrease in thickness of the graded film is related to the fractional rise in the negative charge density of the de-doped PEDT:PSS complex. All the films investigated have similar surface wetting properties, as indicated by dynamic contact angle experiments (H₂O advancing angle 70–80°, compared to 30–45° on bare ITO (ref. 16)). We have also measured the built-in voltage in completed LED structures using electroabsorption spectroscopy, in which the d.c. bias required to null the fundamental response of the electroabsorption signal is determined^{17,18}. There is an increase

in the built-in voltage of 0.18 ± 0.05 V when the uniform interlayer is replaced by the graded interlayer.

Together these results indicate that a polymeric gradient-doped profile can be sustained. The LED characteristics to be described next can be primarily related to the electronic structure of this interlayer.

Diagnostic LEDs were fabricated from ITO anodes modified with an 80-Å (PEDT:PSS/PXT)₆ injection interlayer. In device 1, this was assembled from p^{++} -PEDT:PSS solutions; device 2, from i-PEDT:PSS; device 3, from sequentially de-doped PEDT:PSS (3 p^{++} - p -i). For controls on bare ITO, device 4 was fabricated with no interlayer; and device 5 with a 320-Å PEDT:PSS film spin-cast from as-received PEDT:PSS solutions and baked at 70–80 °C for 1 h. An 860-Å film of poly(9,9-dioctylfluorene) (F8) blended with 5 wt% poly(9,9-dioctylfluorene-*alt*-benzothiadiazole) (F8BT, ref. 19) to give a green emission with a centre wavelength (λ_{ctr}) of 550 nm was spin-cast from *p*-xylene, followed by thermal evaporation of the calcium cathode. This polymer system has a substantial Δ_h ($I_p = 5.8$ – 6.0 eV) with respect to ITO, but a small electron-injection barrier with respect to Ca. As a result, the diode current in conventional diode structures is dominated by electrons, so that any improvement to hole injection or electron confinement will increase their electroluminescence efficiency.

The current–voltage (I – V) and efficiency–voltage (η – V) characteristics of these LEDs are plotted in Fig. 3. Device 4 with no hole-injection layer shows poorly-behaved I – V and η – V characteristics and a low efficiency of 0.1 cd A^{–1}. Device 5 with the spin-cast PEDT:PSS hole-injection layer shows a fourfold increase in the efficiency to 0.4 cd A^{–1}, but also significant levels of low-voltage leakage current.

In contrast, all the devices (1–3) fabricated with (PEDT:PSS/PXT)₆ show well-behaved I – V characteristics. They exhibit a sharp turn-on in both current (threshold of 1 μ A cm^{–2}) and luminance (threshold of 1 mcd m^{–2}) at 2.5 V, which corresponds to the π – π^* energy gap. The low-voltage leakage current is suppressed by two orders of magnitude. The external forward electroluminescence quantum yield (η_{el} , photons emitted into the viewing hemisphere per electron flow) of these devices increases fiftyfold to 6.5% (22 cd A^{–1}) over 2.8–3.2 V. This gives a luminous efficiency of 20–25 lm W^{–1} (at 10–20 μ A cm^{–2}) immediately after the onset of bipolar injection. Device 3 with the graded PEDT yields a current density more than

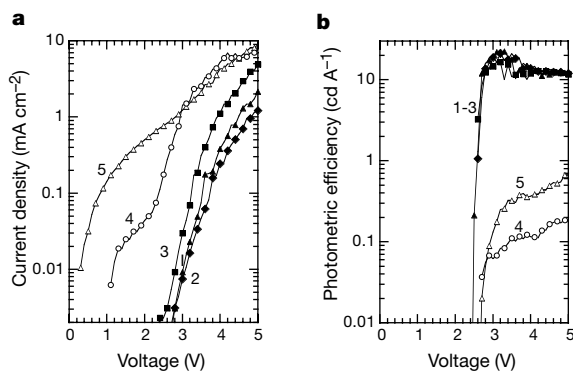


Figure 3 Performance of LEDs with an 860-Å F8-F8BT film, Ca cathode and various ITO anodes. **a**, Current–voltage characteristics. **b**, Photometric efficiency–voltage characteristics. Devices 1–3 have an 80-Å (PEDT:PSS/PXT)₅ injection interlayer: device 1, p⁺⁺-PEDT; device 2, i-PEDT; device 3, p⁺⁺-p⁺-i-PEDT. Control devices: 4, no interlayer; 5, spin-cast PEDT:PSS (320-Å). Diode active area, 1.5 × 2.5 mm. λ_{ctr} , 550 nm.

twice as large as the p⁺⁺-PEDT device 1, and four times that of the i-PEDT device 2. This confirms that appropriate grading can improve hole-injection into the LEP.

LEDs based on the orange-emitting ($\lambda_{\text{ctr}} = 635$ nm) 2,5-dialkoxy-PPV polymer (OC₁C₁₀-PPV, $I_p = 5.2$ –5.4 eV) and the green-emitting ($\lambda_{\text{ctr}} = 560$ nm) [2-alkoxyphenyl-PPV]-co-[2-alkoxyphenyl-5-alkoxy-PPV] polymer ($I_p = 5.4$ –5.6 eV) (chemical structures shown in Fig. 1) have smaller Δ_h than the F8-F8BT LEDs. For these PPV LEDs (ref. 20), grading of the hole-injection interlayer may be less essential. Indeed, we have obtained good diode performances with a (p⁺⁺-PEDT:PSS/PXT)₅(PSS/PXT) injection interlayer that bears a termination barrier layer of the partially-conjugated PPV. For the 20-Å calcium cathode capped by a 1,000-Å Ag layer, this interlayer gives a considerably higher diode current than the graded analogue. We now describe the characteristics of two series of efficient diodes built with this architecture.

LEDs with a 720-Å film of the green-emitting PPV (spin-cast from toluene) turn on at 2.2 V to give a forward η_{el} of 6.0% (20 cd A⁻¹) over 4–6 V and a luminance output of 1,600 cd m⁻² for 8 mA cm⁻² at 5 V (Fig. 4). LEDs with a 620-Å film of the orange-emitting OC₁C₁₀-PPV (spin-cast from 3:7 tetrahydrofuran–toluene) turn on at 1.8 V to give a forward η_{el} of 1.8% (2.6 cd A⁻¹) over 3–5 V and a luminance of 1,000 cd m⁻² for 40 mA cm⁻² at 5 V. These devices show extremely low levels of leakage currents of the order of <1–10 nA cm⁻² at 0.5 V below turn-on. This is critical for large-array passive matrix display applications. These devices also do not show any decline in η_{el} as the LEP film thickness decreases to 650–700 Å, in contrast to the behaviour of earlier generations of polymer LEDs. Therefore low drive voltages can be attained to give high power efficiencies. With further optimization, by cathode engineering²¹ for example, even higher diode current and luminance output could be expected.

To characterize the electron–hole recombination in these remarkably efficient LEDs, we have analysed their external emission pattern by optical modelling²². We summarize the results as follows: (1) electroluminescence arises from in-plane dipoles that give a better surface optical out-coupling than isotropic dipoles. (2) The average position of the emission zone (and hence the recombination zone) in these diodes is near the optimal location for emission through the diode surface. (3) The internal electroluminescence yield is estimated to be 15–20% for the green-PPV diode and 4–5% for the orange-PPV diode. For comparison, the respective free-space photoluminescence yield of the LEPs is 33% and 9%, measured in a N₂-filled integrating sphere on films prepared in the same process run. (4) The γr_{st} product² (where γ is the fraction

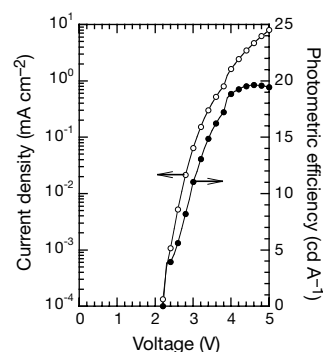


Figure 4 Performance of LEDs with a 720-Å layer of green-emitting PPV (chemical structure in Fig. 1). These devices had Ca/Ag cathodes, and ITO/60-Å-(PEDT:PSS/PXT)₅(PSS/PXT) anodes. Diode active area, 2.0 × 7.5 mm. λ_{ctr} , 560 nm. The peak efficiency of 20 cd A⁻¹ corresponds to 6.0% (photon/carrier) external quantum yield in the forward hemisphere and an internal electroluminescence quantum yield of 15–20%. Free-space photoluminescence yield of this LEP is 33%.

of injected carriers that forms an exciton, and r_{st} is the fraction of excitons formed in the singlet state) for both these diodes are 0.35–0.45. To extract this value, photonic structure effects have been taken into account; also, the probabilities for singlet excitons formed by photoexcitation and by electron–hole recombination to reach the energetically-relaxed emissive state were assumed to be similar, as indicated by the lack of structure in the photoluminescence excitation spectra of related PPVs²³.

This γr_{st} product represents the efficiency for the generation of singlet excitons by electrical injection. Because $\gamma \leq 1$, $r_{\text{st}} \geq 0.35$ –0.45 in these devices. Their singlet-to-triplet exciton formation ratio is therefore considerably larger than the 1:3 spin statistics. This agrees with a study of OC₁C₁₀-PPV LEDs¹, but contrasts with another study²⁴ of aluminium tris(8-hydroxyquinolate) LEDs that gave a singlet-to-triplet ratio of 1:3. The results suggest that the device physics of conjugated polymers differs substantially from that of molecular chelates. Both the surface out-coupling and singlet-to-triplet factors favour a higher electroluminescence efficiency for the polymer devices.

These results may provide useful hints for alternative routes to achieve high device efficiencies. Improved adhesion of the polymer film to the buffer interlayer may be important. We consider that these results not only set a new performance standard for organic LEDs, but also herald interesting new possibilities in the science and technology of molecular electronics. □

Methods

Compensation of PEDT:PSS

10×10^{-3} mol l⁻¹ (by repeat unit) PEDT:PSS solutions ($\gamma = 0.3$) (Bayer) were reduced with measured aliquots of N₂H₄ hydrate (Aldrich) at 80 °C for 2 h and monitored by electronic spectroscopy. Reduction does not proceed appreciably at room temperature. 250 mol equiv. was required to reach $\gamma = 0.02$. All the final solutions contained excess N₂H₄, and were stable against flocculation over several months at room temperature. The PEDT could be saturation-doped again by persulphate oxidation, thereby confirming the absence of chain degradation. Unlike polyanilines, the doping level of PEDT is more stable to pH changes.

Fabrication of the hole-injection interlayers and devices

In a typical polyelectrolyte assembly, the ITO substrates derivatized with trimethylammonioethyl groups⁷ were immersed in the appropriate 0.1 wt% ($\sim 10 \times 10^{-3}$ mol l⁻¹ (by repeat unit)) polyelectrolyte solution at 22 °C for 10 min, rinse-dumped twice in fresh Millipore water, and then immersed in the counter-polyelectrolyte solution, rinse-dumped, and this cycle repeated with no drying-out in-between. The polyelectrolytes included PXT (the precursor polymer of PPV, Cambridge Display Technology), PSS (Na salt, Aldrich), PEDT:PSS (Bayer) and its compensated forms. The ionic strength was typically set to 10 mM with NH₄Cl, and the pH set at 9–10 with 70 mM aqueous NH₃. Upon completion of the assembly, the substrates were baked at 70–80 °C for 2 h under vacuum or dry N₂. This eliminates tetrahydrothiophene from PXT to form covalent

sulphonate crosslinks further immobilizing the dopant polyions (evidenced by the emergence of S—O—C infrared vibrations²⁵ at 1,009 cm⁻¹ and 833 cm⁻¹). Fabrication of 1.5 × 2.5 mm or 2.0 × 7.5 mm diodes was subsequently completed in the glove box without exposing to air. Calcium cathodes were evaporated at a base pressure < 10⁻⁶ mbar.

Characterization

IVL characteristics were collected on encapsulated devices or under vacuum using digital multimeters and a calibrated Si photodiode. Ionization potentials were estimated from the onset of the oxidation peak in cyclic voltammetry experiments. X-ray photoelectron spectroscopy was performed on a VG-220i-XL system with Al K α X-rays. Spectroscopic ellipsometry was recorded on a UVISEL phase-modulated ellipsometer in the polarizer—compensator—sample—analyser configuration with 56.8° incidence angle. Atomic force microscopy was conducted on a Dimension 3100 microscope. Contact angle measurements were measured by a Krauss G20 drop shape analyser. Electroabsorption spectroscopy was performed on ITO/interlayer/green-emitting-PPV/Ca/Ag diodes.

Received 11 November 1999; accepted 1 February 2000.

- Cao, Y., Parker, I. D., Yu, G., Zhang, C. & Heeger, A. J. Improved quantum efficiency for electroluminescence in semiconducting polymers. *Nature* **397**, 414–417 (1999).
- Friend, R. H. *et al.* Electroluminescence in conjugated polymers. *Nature* **397**, 121–128 (1999).
- Karg, S., Scott, J. C., Salem, J. R. & Angelopoulos, M. Increased brightness and lifetime of polymer light-emitting diodes with polyaniline anodes. *Synth. Met.* **80**, 111–117 (1996).
- Carter, J. C. *et al.* Operating stability of light-emitting polymer diodes based on poly(p-phenylene vinylene). *Appl. Phys. Lett.* **71**, 34–36 (1997).
- Decher, G. Fuzzy nanoassemblies: toward layered polymeric multicomposites. *Science* **277**, 1232–1237 (1997).
- Onitsuka, O., Fou, A. C., Ferreira, M., Hsieh, B. R. & Rubner, M. F. Enhancement of light emitting diodes based on self-assembled heterostructures of poly(p-phenylene vinylene). *J. Appl. Phys.* **80**, 4067–4071 (1996).
- Ho, P. K. H., Gränström, M., Friend, R. H. & Greenham, N. C. Ultrathin self-assembled layers at the ITO interface to control charge injection and electroluminescence efficiency in polymer light-emitting diodes. *Adv. Mater.* **10**, 769–774 (1998).
- Tai, K., Yang, L., Wang, Y. H., Wynn, J. D. & Cho, A. Y. Drastic reduction of series resistance in doped semiconductor distributed Bragg reflectors for surface-emitting lasers. *Appl. Phys. Lett.* **56**, 2496–2498 (1990).
- Arkhipov, V. I., Emelianova, E. V., Tak, Y. H. & Bässler, H. Charge injection into light-emitting diodes: theory and experiment. *J. Appl. Phys.* **84**, 848–856 (1998).
- Yoo, D., Shiratori, S. S. & Rubner, M. F. Controlling bilayer composition and surface wettability of sequentially adsorbed multilayers of weak polyelectrolytes. *Macromolecules* **31**, 4309–4318 (1998).
- Lösche, M., Schmitt, J., Decher, G., Bouwman, W. G. & Kjaer, K. Detailed structure of molecularly thin polyelectrolyte multilayer films on solid substrates as revealed by neutron reflectometry. *Macromolecules* **31**, 8893–8906 (1998).
- Dietrich, M., Heinze, J., Heywang, G. & Jonas, F. Electrochemical and spectroscopic characterisation of polyalkylenedioxythiophenes. *J. Electroanal. Chem.* **369**, 87–92 (1994).
- Lögdlund, M., Lazzaroni, R., Stafström, S., Salaneck, W. R. & Brédas, J.-L. Direct observation of charge-induced pi-electronic structural changes in a conjugated polymer. *Phys. Rev. Lett.* **63**, 1841–1844 (1989).
- Xing, K. Z., Fahlman, M., Chen, X. W., Ingañá, O. & Salaneck, W. R. The electronic structure of poly(3,4-ethylenedioxythiophene) studied by xps and ups. *Synth. Met.* **89**, 161–165 (1997).
- van Slyke, S. A., Chen, C. H. & Tang, C. W. Organic electroluminescent devices with improved stability. *Appl. Phys. Lett.* **69**, 2160–2162 (1996).
- Kim, J. S., Friend, R. H. & Cacialli, F. Surface energy and polarity of treated indium-tin-oxide anodes for polymer light-emitting diodes studied by contact-angle measurements. *J. Appl. Phys.* **86**, 2774–2778 (1999).
- Campbell, I. H., Hagler, T. W. & Smith, D. L. Direct measurement of conjugated polymer electronic excitation energies using metal/polymer/metal structures. *Phys. Rev. Lett.* **76**, 1900–1903 (1996).
- Brown, T. M. *et al.* Built-in field electroabsorption spectroscopy of polymer light-emitting diodes incorporating a doped poly(3,4-ethylenedioxythiophene) hole injection layer. *Appl. Phys. Lett.* **75**, 1679–1681 (1999).
- He, Y., Gong, S., Hattori, R. & Kanicki, J. High performance organic polymer light-emitting heterostructure devices. *Appl. Phys. Lett.* **74**, 2265–2267 (1999).
- Spreitzer, H. *et al.* Soluble phenyl-substituted PPVs - new materials for highly efficient polymer LEDs. *Adv. Mater.* **10**, 1340–1343 (1998).
- Hung, L. S., Tang, C. W. & Mason, M. G. Enhanced electron injection in organic electroluminescence devices using an Al/LiF electrode. *Appl. Phys. Lett.* **70**, 152–154 (1997).
- Kim, J. S., Ho, P. K. H., Greenham, N. C. & Friend, R. H. Electroluminescence emission pattern of organic light-emitting diodes: implications for device efficiency calculations. *J. Appl. Phys.* (in the press).
- Harrison, N. T., Hayes, G. R., Phillips, R. T. & Friend, R. H. Singlet intrachain exciton generation and decay in poly(p-phenylenevinylene). *Phys. Rev. Lett.* **77**, 1881–1884 (1996).
- Baldo, M. A., O'Brien, D. F., Thompson, M. E. & Forrest, S. R. Excitonic singlet-triplet ratio in a semiconducting organic thin film. *Phys. Rev. B* **60**, 14422–14428 (1999).
- Colthup, N. B., Daly, L. H. & Wiberley, S. E. *Introduction to Infrared and Raman Spectroscopy* (Academic, New York, 1964).

Acknowledgements

We thank I. Grizzi, D. J. Lacey and E. P. Woo for support; J.-W. Cai for X-ray photoelectron spectroscopy; and A. Gerhard for electroabsorption measurements. P.K.H.H. is on leave from the National University of Singapore and thanks St John's College and IMRE for funding. This work was supported in part by the Engineering and Physical Sciences Research Council.

Correspondence and requests for materials should be addressed to R.H.F. (e-mail: rhf10@cam.ac.uk).

Effect of climate change relative to ozone depletion on UV exposure in subarctic lakes

Reinhard Pienitz & Warwick F. Vincent

Centre d'Études Nordiques, Université Laval, Québec, Canada G1K 7P4

The effect of stratospheric ozone depletion on increases in ambient levels of solar ultraviolet (UV) radiation in high-latitude regions¹ has raised concerns about the response of northern ecosystems to environmental change. The concentration of coloured dissolved organic material, which is derived from terrestrial vegetation and acts as a screen for ultraviolet radiation, is low in high-latitude lakes². The underwater light environment in these lakes is therefore likely to be sensitive to small variations in the supply of this material, in addition to the effects of ozone depletion^{2–5}. Here we use fossil diatom assemblages in combination with bio-optical models to estimate the magnitude of past variations in the underwater light regime of a lake at the boreal tree line. We find large shifts in underwater UV-B, UV-A and photosynthetically available radiation associated with changes in the input of coloured dissolved organic material into subarctic lakes during the Holocene. The inferred changes in biological exposure to UV radiation were at least two orders of magnitude greater than those associated with moderate (30%) ozone depletion. Our findings indicate that freshwater ecosystems at present located across vegetation gradients will experience significant shifts in underwater spectral irradiance through the effects of climate change on catchment vegetation and the export of coloured dissolved organic material.

To address the potential effect of climate change relative to ozone depletion on northern freshwater ecosystems, we combined palaeolimnological analyses with bio-optical models based on present-day conditions. Specifically, we estimated past underwater light conditions from concentrations of dissolved organic carbon (DOC, a correlate of coloured dissolved organic material, CDOM^{2–5}) that were inferred from fossil diatom assemblages preserved in Holocene sedimentary deposits⁶. Strong statistical relationships exist between diatom community structure and organic carbon concentrations^{7–9}, which have been used to infer past DOC concentrations from fossil diatom records^{6,10}. The sediments were sampled from Queen's Lake (unofficial name; latitude 64°07'N, longitude 110°34'W; Fig. 1) in the central Northwest Territories, Canada. The lake basin is located on Precambrian granites of the Canadian Shield ~25 km north of the mapped limit of the central Canadian tree line, and lies on the boundary between the high subarctic and low Arctic ecoclimate

Table 1 Inferred DOC changes in two subarctic lakes during the Holocene

Period	Dates (kyr BP)	Queen's Lake		Toronto Lake
		DOC (mg l ⁻¹)	DV (10 ⁷ g ⁻¹)	DOC (mg l ⁻¹)
Post	<3	1.4 (0.2)	5.33 (1.5)	2.8 (0.2)
Mid	3–5	5.6 (0.2)	21.09 (3.3)	4.2 (0.1)
Pre	>5	1.8 (0.3)	8.87 (2.1)	n.d.

Each value is the mean (± s.e.) for the period after (Post), during (Mid) and before (Pre) the mid-Holocene warm interval for the central Canadian treeline region. A two-way ANOVA analysis (general linear model, unequal design, no interaction terms) of the DOC data for the two lakes showed that there are statistically significant effects of period ($F = 57.1$; $P < 0.001$) but no significant differences between lakes ($F = 1.3$; $P = 0.26$). A subsequent Tukey multiple comparison test showed a significant increase between Pre and Mid ($P < 0.001$), a significant decrease between Mid and Post ($P < 0.001$), but no significant difference between Pre and Post ($P = 0.112$) periods. Diatom valve concentrations were higher during the mid-Holocene warm interval ($P < 0.05$; Tukey multiple comparison test) relative to Pre and Post periods, which were not significantly different from each other ($P = 0.73$). DV, diatom valves; n.d., not determined because the diatom community composition was outside the bounds of the training set.

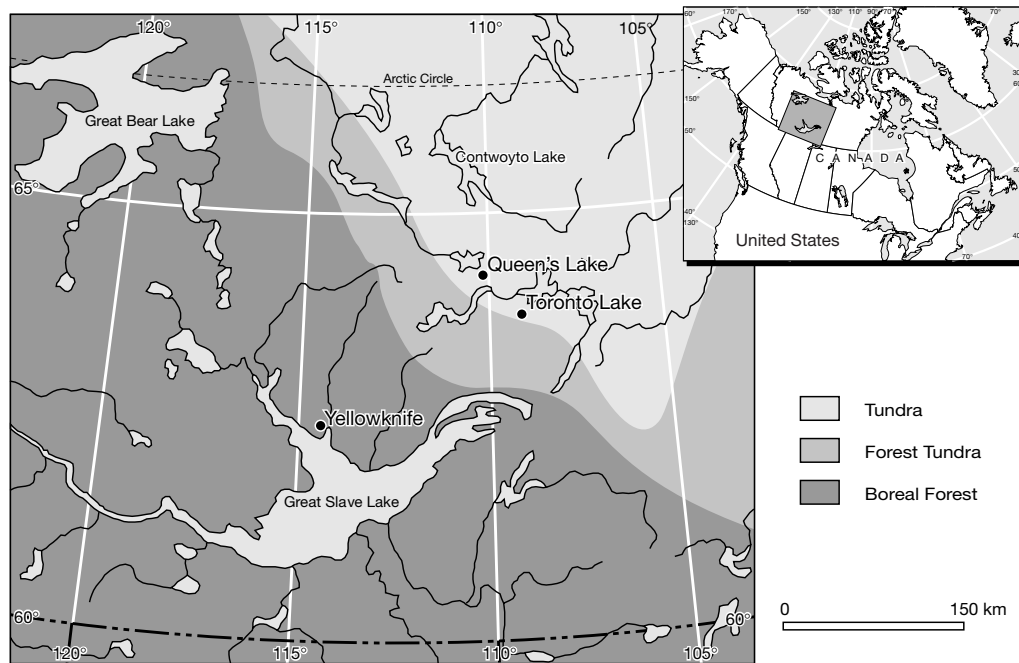


Figure 1 Location of study sites near the northern tree line in central Canada. The catchments of both lakes are typified by rolling terrain, composed mainly of Proterozoic intrusive granites (granodiorites). Cryic regosols and organic soils are found on discontinuous glacial overburden underlain by permafrost. Queen's Lake (480 m above sea level, a.s.l.; area, 50 ha; depth, 3.5 m) is a closed basin, whereas Toronto Lake

(414 m a.s.l.; area, 10 ha; depth, 6.75 m) is hydrologically open and drains an extensive headwater catchment (~2,900 ha). Both lakes are located in *Betula glandulosa* tundra, with a few patches of *Picea mariana* Krummholz in sheltered locations. Further details on regional geology, climate and soils are given in refs. 11, 12, 14.

regions. The tree line is defined as the zone of transition between closed continuous forest and tundra. Details on diatom stratigraphies and sediment core chronology are presented and discussed elsewhere^{6,11}.

Diatom community structure and diatom-inferred DOC concentrations show three distinct and abrupt successional changes over the history of Queen's Lake (Fig. 2a, b). A major shift occurred between ~5 and ~3 kyr before present (BP) in the proportion of periphytic versus planktonic diatom species, corresponding to peak abundance (>70%) of periphytic taxa and a strong decline in the relative abundance of planktonic species. This period also corresponded to a significant increase in the total diatom population, as suggested by peak concentrations of diatom valves per unit sediment (Table 1); the shift towards periphyton dominance at this time was therefore not simply a result of reduced plankton concentrations. The profile of diatom-inferred DOC indicates highest levels between 5 and 3 kyr BP, showing a threefold increase in concentrations when compared with the periods preceding and following this mid-Holocene warm interval (Fig. 2c). The diatom-based reconstruction then indicates a sharp decline (~88%) in DOC concentrations over the most recent 3,000 years.

These biological and chemical shifts correspond with other evidence of major climatic and other environmental changes in central Canada following local deglaciation around 9 kyr BP; this period was characterized by the advance and subsequent retreat of the boreal tree line^{11–13}, accompanied by profound limnological and hydrological changes^{6,11,14}, in response to shifts in the mean summer position of the Arctic frontal zone¹⁵. Terrestrial vegetation abruptly shifted from dwarf shrub tundra to *Picea mariana* forest-tundra about 5 kyr BP, as the frontal zone moved northwards. Minor local fluctuations in treeline position or forest density occurred during the period of maximum forest cover (forest-tundra) between 5 and 3 kyr BP, followed by a return to the current dwarf shrub tundra vegetation after 3 kyr BP¹¹.

DOC concentration, and its correlates water colour and CDOM, decline abruptly with decreasing forest cover across the northern

tree line^{3,16,17}. As found for shifts in the alpine tree line¹⁸, the Holocene vegetational changes in central Canada are likely to have been accompanied by significant variations in the export of CDOM (mostly humic and fulvic acids) from the catchments, which in turn would influence the underwater UV radiation in their receiving waters. Our data from Queen's Lake are consistent with such effects, and suggest that the highest sustained lakewater DOC concentrations coincided with the period of maximum tree-line advance and/or highest forest cover density during the mid-Holocene (about 5,000–3,000 years ago). These limnological changes are not unique to the Queen's Lake site, as similar diatom and DOC records have been obtained from another treeline lake (Toronto Lake, Fig. 1; Table 1)^{6,11}, reflecting the regional character of climatic, vegetational and limnological changes that occurred near the northern tree line in central subarctic Canada during the Holocene^{12–14}.

The large and rapid changes in DOC imply that Queen's Lake also experienced significant shifts in the underwater optical environment over the past 6,000 years, as its correlate CDOM is the main attenuator of UV radiation and photosynthetically available radiation (PAR) in oligotrophic lakes of northern regions, including the boreal forest^{2,5}. Application of our bio-optical models (which have been derived from measurements in high-latitude waters, including lakes across the North American tree line) showed that the inferred DOC shifts were equivalent to a decrease by two orders of magnitude in biologically effective UV exposure between 6 kyr BP and the mid-Holocene vegetation maximum; the most recent 3,000 years would have been characterized by a >50-fold increase to present-day values (Fig. 2d).

These effects were consistently large for two different biological weighting functions; the first for DNA damage (particularly sensitive to UV-B) and the second for UV-photoinhibition of algal photosynthesis (sensitive to UV-A plus UV-B). The inferred DOC shifts also imply substantial changes in underwater PAR over the course of the Holocene, with minimum light availability for photosynthesis during the vegetation maximum. This PAR effect

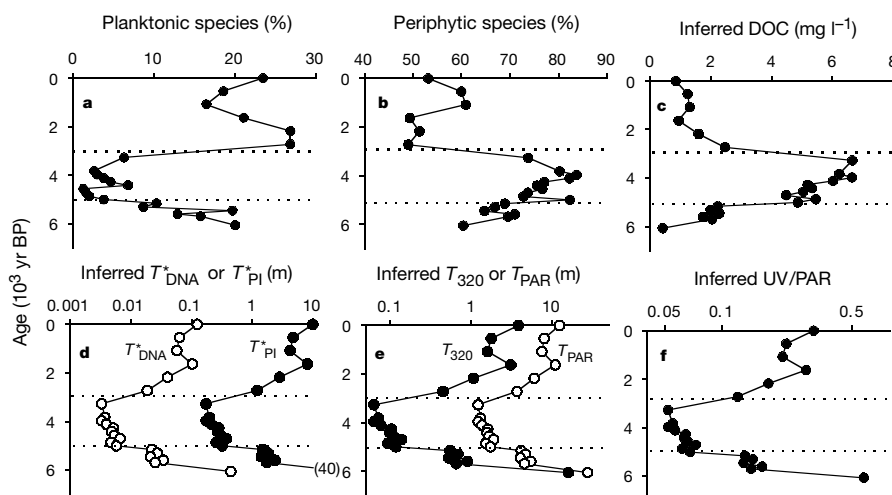


Figure 2 Changes in diatom community structure and inferred variables in Queen's Lake. The diatom data are expressed as the percentage of the total number of valves in each sample associated with planktonic or benthic taxa. Diatom species data were used to infer DOC (mg l^{-1}), biologically weighted UV exposure (T^*_{PI} or T^*_{DNA}) and underwater spectral

balance (water column transparency for 320-nm UV radiation (T_{320}), PAR (T_{PAR}), and the ratio between the two (UV/PAR)) over the past 6,000 years. The dotted lines show the period of mid-Holocene maximum forest cover.

was less than for UV radiation (Fig. 2e), and the inferred water column ratio of UV radiation (at 320 nm) to PAR dropped by an order of magnitude between 6 to 3.5 kyr BP, followed by a >5-fold rise over the last 3,000 years (Fig. 2f). We note that the vegetation maximum also corresponds to the period of large increase in periphytic diatom taxa relative to planktonic diatoms. The planktonic species may have been more sensitive to PAR-limitation in the less transparent waters during this period of high DOC. Conversely, the periphyton may have been less UV-inhibited at this time, given the substantially reduced UV/PAR ratio. Elsewhere in Canada, UV radiation has been shown to have a pronounced impact on periphytic algal communities in shallow lakes¹⁹, and decreased UV exposure in shallow littoral habitats may be one reason for the predominance of periphytic diatom taxa during periods of high DOC, lowest biologically effective UV exposure and lowest water transparency.

The biologically weighted UV transparency parameter (T^* , Fig. 2) allows a comparison of the effects of DOC change with those associated with stratospheric ozone depletion on a common, biologically relevant, scale. For this analysis we recomputed T^* for a 30% decline (over the period late 1970s to early 1990s there was a 10% depletion of ozone over the Arctic¹, with more recent evidence of up to 40% depletion in early spring²⁰) in stratospheric ozone, from 330 to 230 Dobson units (DU). Under DOC conditions of 1 mg Cl^{-1} (the approximate value inferred for Queen's Lake for the past 1,000 years), this would result in a 24% increase in biological UV exposure weighted for DNA effects, and a 0.9% increase weighted for photoinhibition effects. These increases are small relative to those calculated from the changes in inferred DOC over the past 6,000 years: from the vegetation maximum to the present, the estimate of T^* for DNA increases by 5,651% while T^* for photoinhibition increases by 3,658%. The latter is especially striking relative to the ozone effect. This is because, in contrast to ozone, CDOM controls UV-A as well as UV-B attenuation, and UV-A is a significant control on photoinhibition under natural spectral regimes²¹.

Numerical integration of weighted UV radiation from the surface to the current maximum depth of the lake (3.5 m) confirmed the much greater effect of changes in dissolved colour. For the more ozone-sensitive DNA effect, this calculation gave a total water

column weighted UV that was 118% higher under conditions of 230 DU but 1,079% higher under the inferred DOC reduction; the equivalent increases for UV-photoinhibition were 7% (reduced ozone) and 1,456% (reduced colour). We note, however, that the ozone depletion effect is taking place over a timescale of decades, while the vegetation-related changes are likely to have occurred over a timescale of 100 years and longer.

The T^* estimates based on DOC must be considered a lower bound to variability in underwater UV exposure; this exposure is also influenced by other variables, including biotic and abiotic particle concentrations, and climate-related changes in cloud, lake-ice and snow cover. Changes in lake depth and mixing regimes³, as well as threshold, recovery and other nonlinear effects of UV exposure on biological processes, will also contribute to the variability in response. However, these calculations provide evidence for significant variations in biological UV exposure during the Holocene that far exceed those associated with current levels of ozone depletion, and they draw attention to the sensitivity of these freshwater ecosystems to future shifts in vegetation and hydrology associated with global climate change.

Global warming scenarios predict a seasonal or annual warming of the order of 1–4 °C in the treeline zone of Canada in the next century²². Examination of past warm episodes, such as the mid-Holocene climate optimum, may provide the best available analogue scenarios with respect to the anticipated climate warming and its associated limnological and hydrological effects. The potential for a large northward displacement of the forest boundary in central Canada²³ and the effect of such a displacement on climate²⁴ makes it important to understand the potential response of this region and its abundant freshwater ecosystems to climate change. It is likely that warming would lead to relatively rapid infilling of forest in the forest-tundra zone and associated changes in the physical and chemical conditions of freshwater lakes⁶, with ecologically significant effects within ~100–150 years (refs 11, 13). The Holocene chemical and bio-optical changes inferred for Queen's Lake are likely to be representative of many lakes across the circumpolar treeline region, as well as other ecotonal boundaries such as the alpine tree line, and this study provides some insights into the likely response of freshwater aquatic ecosystems at the tree line to future warming.

The history of palaeo-optical conditions inferred from Queen's Lake implies that aquatic biota of lakes in the boreal treeline zone may have experienced their greatest UV exposure immediately following Holocene deglaciation (before the development of CDOM sources in the catchment) and during subsequent episodes of reduced CDOM inputs; such episodes occurred when climate cooling reduced forest canopy and shifted the tree line, ultimately causing the catchment basin to lose its forest cover (for example, the Neoglacial period of colder and drier climate conditions since ~ 3 kyr BP^{6,14}). Our study also shows that global warming effects in lakes located near the present-day northern boundary of the boreal forest may be very different from those observed in Precambrian Shield lakes to the south, where some lakes lost 50% of their DOC over a 20-year period during the 1970s and 1980s due to acidification and climate warming^{4,25}. UV penetration in dilute lakes that are presently within catchments dominated by tundra and forest-tundra may be significantly reduced whenever climate warming induces northward shifts of tree line or increases in the density of forest cover, leading to increased terrestrial CDOM inputs. The reduced UV-inhibition of biological processes is partially offset, however, by reduced underwater light for photosynthesis and decreased photochemical degradation of DOC into substrates that are more available for bacterial production. Also, changes in wetland patterns in the catchment associated with warmer conditions are likely to modify DOC concentrations in the recipient lakes²⁶. Lakes with the highest catchment area to lake volume ratios are likely to be the most affected by these vegetational changes.

Our palaeo-optical analysis of Queen's Lake shows that the UV climate of lakes at, and near, the tree line may have been changed more radically by climate change than by moderate ozone depletion. The Holocene record provides a window into future changes in this vast circumpolar biome, and implies that current warming trends in the subarctic may lead to significant shifts in underwater spectral balance, PAR availability and UV radiation exposure for aquatic organisms. □

Methods

Palaeolimnological analysis

Queen's Lake and Toronto Lake were sampled in the region of maximum depth where the sediments provided an integration of material from the littoral zone (containing periphyton) as well as the offshore limnetic zone (containing phytoplankton). Sediment samples were taken at 2.5-cm intervals from the AMS ¹⁴C-dated sediment cores; diatoms were extracted, isolated, identified and quantified following standard procedures^{6,11}.

Past DOC concentrations in Queen's Lake were reconstructed from the relative abundances of fossil diatoms using a weighted-averaging partial least squares (WA-PLS) transfer function for DOC developed from diatoms in dilute, oligotrophic lakes from the study region in the central Northwest Territories⁶, and therefore appropriate for Queen's Lake. Development of the diatom-DOC model involved a two-step analytical approach⁶. First, responses of modern diatoms to DOC in a set of lakes equally distributed across the northern tree line (DOC gradient, 1.6–9.1 mg l⁻¹; $n = 22$) were modelled in a weighted-averaging (WA) regression. Second, these modelled responses were used to infer past DOC concentrations from the composition of fossil diatom assemblages using WA calibration. The root-mean-squared error of prediction in the estimates of inferred DOC is 1.3 mg C l⁻¹ with a jack-knifed r^2 of 0.61 for the diatom-DOC model as calculated by WAPLS version 1.1 (S. Juggins and C.J.F. ter Braak). Performance measures (program ANALOG; J. M. Line and H. J. B. Birks) of the diatom inference model indicated that no analogue situations with modern diatom communities did not occur in the Queen's Lake core, but within the period preceding the mid-Holocene warm interval in the Toronto Lake core⁶. Data for planktonic and periphytic diatoms and diatom-inferred DOC were based on a total of 76 taxa.

Biological UV exposure

Weighted transparency values (T^*) allowed the relative effects of UV attenuation in the atmosphere and water to be assessed on a common, biologically relevant scale²⁷. T^* is defined as $\int 1/K(\lambda)\epsilon(\lambda)E_{\text{rel}}(\lambda)F(\lambda)d\lambda$, where the integral is evaluated over the UV-B plus UV-A wavelength range (280–400 nm). $K(\lambda)$ is the diffuse attenuation coefficient at wavelength λ calculated from statistical relationships with DOC for high-latitude lakes²⁷; $\epsilon(\lambda)$ is the biological weighting factor for DNA damage²⁸ (T_{DNA}) or for UV-photo-inhibition of photosynthesis²¹ (T^*_{PI}) and set on a relative scale ($\epsilon = 1.0$ at 300 nm); $E_{\text{rel}}(\lambda)$ is the normalized surface irradiance at Churchill in subarctic Manitoba ($E_{\text{rel}} = 1.0$ at 400 nm; we used an incident UV radiation spectrum for Churchill (lat. 58° 75' N, long. 94° 07' W) at low zenith angle (36.57°) from the World Ozone and Ultraviolet Radiation

Data Center database, World Meteorological Organization, Toronto); and $F(\lambda)$ is the factor of enhancement in surface radiation flux for a given stratospheric ozone depletion²⁹ (set to 1.0 for 330 DU). T^* values were calculated at 1-nm intervals and then summed from 290 to 400 nm to give a total T^* for UV radiation. T^* is analogous to other oceanographic indices based on transparency ($1/K$), such as critical depth and average water column irradiance.

Received 28 October 1999; accepted 16 February 2000.

1. International Arctic Science Committee *Effects of Increased Ultraviolet Radiation in the Arctic* (Report No. 2, IASC, Oslo, 1995).
2. Laurion, I., Vincent, W. F. & Lean, D. R. S. Underwater ultraviolet radiation: development of spectral models for northern high latitude lakes. *Photochem. Photobiol.* **65**, 107–114 (1997).
3. Vincent, W. F. & Pienitz, R. Sensitivity of high latitude freshwater ecosystems to global change: temperature and solar ultraviolet radiation. *Geosci. Can.* **23**, 231–236 (1996).
4. Schindler, D. W., Curtis, P. J., Parker, B. R. & Stainton, M. P. Consequences of climate warming and lake acidification for UV-B penetration in North American boreal lakes. *Nature* **379**, 705–708 (1996).
5. Williamson, C. E., Stemberger, R. S., Morris, D. P., Frost, T. M. & Paulsen, S. G. Ultraviolet radiation in North American lakes: Attenuation estimates from DOC measurements and implications for plankton communities. *Limnol. Oceanogr.* **41**, 1024–1034 (1996).
6. Pienitz, R., Smol, J. P. & MacDonald, G. M. Paleolimnological reconstruction of Holocene climatic trends from two boreal treeline lakes, Northwest Territories, Canada. *Arct. Antarct. Alp. Res.* **31**, 82–93 (1999).
7. Pienitz, R. & Smol, J. Diatom assemblages and their relationship to environmental variables in lakes from the boreal forest-tundra ecotone near Yellowknife, Northwest Territories, Canada. *Hydrobiologia* **269/270**, 391–404 (1993).
8. Battarbee, R. W., Flower, R. J., Juggins, S., Patrick, S. T. & Stevenson, A. C. The relationship between diatoms and surface water quality in the Høylandet area of Nord-Trøndelag, Norway. *Hydrobiologia* **348**, 69–80 (1997).
9. Fallu, M. -A. & Pienitz, R. Diatomées lacustres de Jamésie-Hudsonie (Québec) et modèle de reconstitution des concentrations de carbone organique dissous. *Écoscience* **6**, 603–620 (1999).
10. Korsman, T., Renberg, I. & Anderson, N. J. A palaeolimnological test of the influence of Norway spruce (*Picea abies*) immigration on lake-water acidity. *Holocene* **4**, 132–140 (1994).
11. MacDonald, G. M., Edwards, T. W. D., Moser, K. A., Pienitz, R. & Smol, J. P. Rapid response of treeline vegetation and lakes to past climate warming. *Nature* **361**, 243–246 (1993).
12. Moser, K. A. & MacDonald, G. M. Holocene vegetation change at treeline north of Yellowknife, Northwest Territories, Canada. *Quat. Res.* **34**, 227–239 (1990).
13. MacDonald, G. M., Szeicz, J. M., Claricoates, J. & Dale, K. A. Response of the central Canadian treeline to recent climatic changes. *Ann. Assoc. Am. Geogr.* **88**, 183–208 (1998).
14. Wolfe, B. B., Edwards, T. W. D., Aravena, R. & MacDonald, G. M. Rapid Holocene hydrologic change along boreal treeline revealed by $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ in organic lake sediments, Northwest Territories, Canada. *J. Paleolimnol.* **15**, 171–181 (1996).
15. Pielke, R. A. & Vidale, P. L. The boreal forest and the polar front. *J. Geophys. Res.* **100**, 25755–25758 (1995).
16. Engstrom, D. R. Influence of vegetation and hydrology on the humus budgets of Labrador lakes. *Can. J. Fish. Aquat. Sci.* **44**, 1306–1314 (1987).
17. Pienitz, R., Smol, J. P. & Lean, D. R. S. Physical and chemical limnology of 24 lakes located between Yellowknife and Contwoyto Lake, Northwest Territories (Canada). *Can. J. Fish. Aquat. Sci.* **54**, 347–358 (1997).
18. Leavitt, P. R., Vinebrooke, R. D., Donald, D. B., Smol, J. P. & Schindler, D. W. Past ultraviolet radiation environments in lakes derived from fossil pigments. *Nature* **388**, 457–459 (1997).
19. Vinebrooke, R. D. & Leavitt, P. R. Effects of ultraviolet radiation on periphyton in an alpine lake. *Limnol. Oceanogr.* **41**, 1035–1040 (1996).
20. Hansen, G. & Chipperfield, M. P. Ozone depletion at the edge of the Arctic polar vortex 1996/1997. *J. Geophys. Res.* **104**, D1, 1837–1845 (1998).
21. Cullen, J. J., Neale, P. J. & Lesser, M. P. Biological weighting function for the inhibition of phytoplankton photosynthesis by ultraviolet radiation. *Science* **258**, 646–650 (1992).
22. Houghton, J. T. et al. (eds) *Climate Change 1995: The Science of Climate Change* (Cambridge Univ. Press, Cambridge, 1996).
23. Smith, T. M., Shugart, H. H., Bonan, G. B. & Smith, J. B. Modeling the potential response of vegetation to global climate change. *Adv. Ecol. Res.* **22**, 93–116 (1992).
24. Foley, J. A., Kutzbach, J. E., Coe, M. T. & Levis, S. Feedbacks between climate and boreal forests during the Holocene epoch. *Nature* **371**, 52–54 (1994).
25. Yan, N. D., Keller, W., Scully, N. M., Lean, D. R. S. & Dillon, P. J. Increased UV-B penetration in a lake owing to drought-induced acidification. *Nature* **381**, 141–143 (1996).
26. Urban, N. R., Bayley, S. E. & Eisenreich, S. J. Export of dissolved organic carbon and acidity from peatlands. *Wat. Resour. Res.* **25**, 1619–1628 (1988).
27. Vincent, W. F., Laurion, I. & Pienitz, R. Arctic and Antarctic lakes as optical indicators of global change. *Ann. Glaciol.* **27**, 691–696 (1998).
28. Setlow, R. B. The wavelengths in sunlight effective in producing skin cancer: a theoretical analysis. *Proc. Natl. Acad. Sci. USA* **71**, 3363–3366 (1974).
29. Frederick, J. E. & Snell, H. E. Ultraviolet radiation levels during the Antarctic spring. *Science* **241**, 438–440 (1988).

Acknowledgements

We thank G. M. MacDonald for providing the sediment cores, and J. J. Cullen, J. A. E. Gibson, C. Lovejoy and D. W. Schindler for their comments on the manuscript. This work was supported by Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (Québec), Natural Sciences and Engineering Research Council of Canada and Centre d'Études Nordiques.

Correspondence and requests for materials should be addressed to R.P. (e-mail: reinhard.pienitz@cen.ulaval.ca) or W.F.V. (e-mail: warwick.vincent@bio.ulaval.ca).

Hf–Nd isotope evidence for a transient dynamic regime in the early terrestrial mantle

Francis Albarède*, Janne Blichert-Toft*, Jeffrey D. Vervoort†, James D. Gleason* & Minik Rosing‡

* Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon cedex 7, France

† University of Arizona, Tucson, Arizona 85721, USA

‡ Geologisk Museum, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark

Modern basalts have seemingly lost all 'memory' of the primitive Earth's mantle except for an ambiguous isotopic signal observed in some rare gases^{1,2}. Although the Earth is expected to have reached a thermal steady state within several hundred million years (refs 3, 4) of accretion, it is not known how and when the initial chemical fractionations left over from planetary accretion (and perhaps a stage involving a magma ocean) were overshadowed by fractionations imposed by modern-style geodynamics. Because of the lack of samples older than 4 Gyr, this early dynamic regime of the Earth is poorly understood. Here we compare published Hf–Nd isotope data on supracrustals from Isua, Greenland, with similar data on lunar rocks and the SNC (martian) meteorites, and show that, about 3.8 Gyr ago, the geochemical signature of the Archaean mantle was partly inherited from the initial differentiation of the Earth. The observed features seem to indicate that the planet at that time was still losing a substantial amount of primordial heat. The survival of remnants from an early layering in the modern deep mantle may account for some unexplained seismological, thermal and geochemical characteristics of the Earth as observed today.

The suspicion of large-scale element mobility inspired a long-standing controversy on the interpretation of the early Archaean Nd isotope record with ¹⁴³Nd/¹⁴⁴Nd values far different from chondrites⁵. Some authors see in these data a strong indication of a transient stage between the early fractionation of the Earth and its modern convection regime, and discuss various processes of crust extraction and early terrestrial differentiation that may account for this^{6–10}. Others emphasize the mobility of Sm and Nd in these high-grade metamorphic rocks, and interpret the wide variations in initial ¹⁴³Nd/¹⁴⁴Nd ratios to be artefacts due to later metamorphic disturbances^{10–13}. Here we reassess the degree of depletion of the terrestrial mantle at the time of Isua rock formation in the light of new Hf isotope data on Mars and the young Earth published elsewhere^{14,15}.

Lu/Hf and Sm/Nd fractionation by terrestrial magmatic processes are normally strongly correlated, but this is not always the case for fractionation caused by metamorphic disturbances. Such a decoupling is probably due to complexing anions, such as sulphate and fluoride, being present in variable concentrations, and to the chief budget of Hf in a rock residing largely in resistant zircons, whereas the Sm–Nd–Lu budget resides in various silicate minerals. In order to address the question of whether some geochemical imprint of a primitive terrestrial mantle is observable, two steps are required: (1) the identification of a reliable criterion of elemental mobility in ancient rocks; and (2) a common representation of modern and ancient samples that may be used to substantiate a changing chemical fractionation of the mantle.

It is well recognized that the mafic igneous rocks at Isua may have been an open system during metamorphism^{11,12}. The relative mobility of Sm, Nd, Lu and Hf in these rocks has been extensively discussed elsewhere¹⁴ with reference to modern equivalents. All modern oceanic basalts form a regular trend in the Lu/Hf versus Sm/Nd plot (Fig. 1) that can be used to filter out samples affected by

metamorphic disturbances. Among the Sm–Nd and Lu–Hf data on ~3.8-Gyr-old supracrustals (metabasalts and metasediments) from Isua, we excluded all the samples that do not fall on this trend. In particular, all the mafic rocks from the 'garbenschiefer' unit, in which Hf and rare-earth elements are unusually depleted, do not lie on this trend, and have been excluded. When perturbed, most Isua mafic rocks seem to have suffered loss rather than gain of Sm, Nd, Lu and Hf, so that the samples are unlikely to have been contaminated by exotic material. In addition, mafic rocks have Sm/Nd and Lu/Hf ratios close to chondritic values, and therefore the calculation of initial ϵ values is relatively age-insensitive. For granites, felsic gneisses and clastic sediments, we assume, as have previous workers¹³, that the crustal history of these rocks is short enough that their initial ¹⁴³Nd/¹⁴⁴Nd and ¹⁷⁶Hf/¹⁷⁷Hf ratios can be used to provide minimum values for the mantle source of their igneous precursor.

In order to compare the geochemical source characteristics of modern and ancient mantle-derived rocks, we have used the representation of Nyquist *et al.*¹⁶ for lunar rocks. First, the Nd and Hf initial isotopic compositions of the magmatic rocks are calculated at the time of their crystallization. Second, the apparent parent/daughter ratio of the mantle source of the rocks is obtained by subtracting the planetary (bulk silicate Earth, BSE) from the initial isotopic composition and dividing the resulting isotopic ingrowth by their respective decay constants. We show in Fig. 2 the parent/daughter ratios normalized in this way to the chondritic values. The assumptions used so far are that (1) the age of the rock is known, (2) the rock behaved as a closed system, and (3) the BSE isotopic compositions of Nd and Hf 4.56 Gyr ago are known.

To contrast with the early Archaean terrestrial samples, we also show (Fig. 2) data for shergottites (martian basalts, gabbros, and peridotites)¹⁵ and lunar basalts^{17,18}, and show the field of modern terrestrial basalts for comparison purposes. The shergottites have isotopic signatures representative of both martian depleted mantle and ancient enriched basaltic crust¹⁵. The prominent ¹⁴²Nd and ¹⁸²W isotopic anomalies generated by extinct radionuclides in both martian and lunar samples indicate that basaltic material was extracted very early, implying that the respective mantles of these planets were either not vigorously convecting after they formed^{19,20} or that convection produced distinct reservoirs isolated from the beginning. The large apparent spread in Sm/Nd and Lu/Hf ratios of the samples from Mars (shergottites) and the Moon also argue for a heterogeneous mantle, in stark contrast to the moderate spread of samples derived from either the early Archaean or present-day terrestrial mantle.

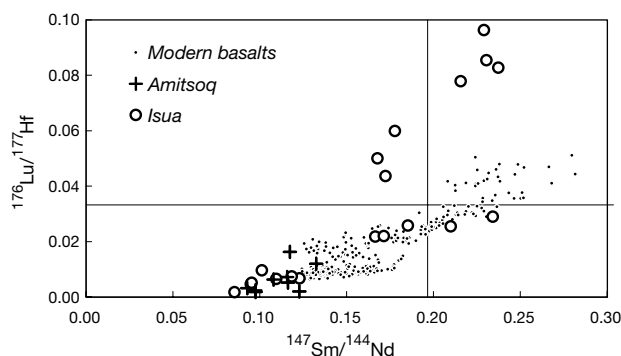


Figure 1 Geochemical screening of the altered Isua metavolcanic samples. Their parent/daughter ratios are compared with those of modern volcanic rocks. Some of the Isua samples, notably the 'garbenschiefer' metabasalts, plot off the field of modern volcanic rocks, indicating that they have been affected by later metamorphism. Only samples with parent/daughter ratios consistent with those of modern volcanic rocks are used to estimate the isotopic properties of the mantle at 3.8 Gyr.

Figure 2 illustrates that, once samples affected by element mobility have been screened using the criterion discussed above, the mantle source of most Isua samples appears to have had a higher Sm/Nd—and some of them also a higher Lu/Hf ratio—than the modern terrestrial mantle. The early Archaean terrestrial mantle was therefore geochemically more depleted (that is, it had higher time-integrated Sm/Nd and Lu/Hf ratios) than has been predicted by extrapolation from the isotopic evolution inferred for the source of Proterozoic and Phanerozoic basalts^{5,10}. This is consistent with the Hf and Nd isotopic compositions of the Amitsoq gneisses, which have initial ϵ_{Hf} and ϵ_{Nd} values that are dominantly more radiogenic than chondritic^{9,10}. Most of the ~3.7-Gyr-old Amitsoq gneisses¹⁰ also fall outside of the modern mantle array on the high-Sm/Nd side. A 3.45-Gyr-old komatiite sample from Barberton was recently shown to present the same anomalous Lu/Hf fractionation²¹. The pattern emerging from ancient terrestrial basalts is therefore intermediate between that of planets with little or no mantle convection (such as Mars and the Moon, which both kept the memory of their initial differentiation), and that of the modern terrestrial mantle, heavily processed by vigorous convection, and presumably at steady state. (We have assumed the age of the Isua supracrustals to be 3.80 Gyr, based in part on the oldest felsic dykes²² that crosscut the metabasalts. The calculated time-integrated parent/daughter values, however, are largely insensitive to the precise age used, within the limits of ± 50 Myr.)

A highly fractionated terrestrial mantle in the Archaean may simply indicate that the current chondritic reference is inadequate for the Earth^{23,24}, and that our planet formed from an unspecified mixture of meteorites from different classes. However, selecting values of the BSE Sm–Nd and Lu–Hf isotopic parameters at 4.56 Gyr that are different by several ϵ units^{10,21,23} from those established in refs 25 and 26 cannot be done arbitrarily. The coincidence between the initial $^{143}\text{Nd}/^{144}\text{Nd}$ and/or $^{176}\text{Hf}/^{177}\text{Hf}$ values in achondrites and chondrites is certainly not fortuitous.

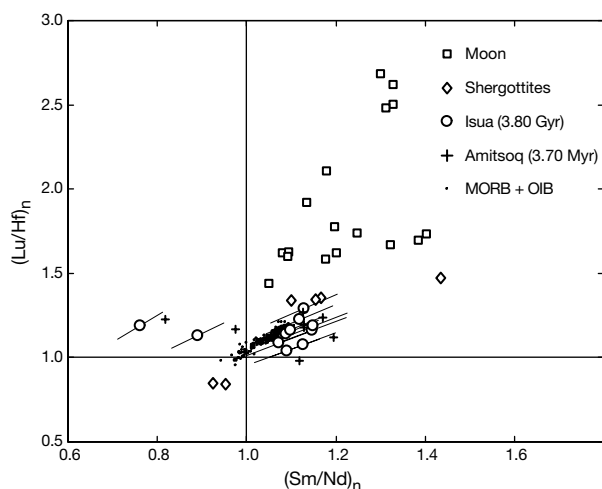


Figure 2 Evidence for strong chemical fractionation in the early Archaean mantle. Shown are time-integrated values of the chondrite-normalized Sm/Nd and Lu/Hf ratios (parent/daughter ratios), estimated for the source of modern basalts (modern literature and unpublished data), SNC meteorites¹⁵, lunar basalts^{17,18}, Isua supracrustals¹⁴ and Amitsoq gneisses¹⁰. In light of recent work^{23,24}, the decay constant of ^{176}Lu has been revised and we now use a value of $1.876 \times 10^{-11} \text{ yr}^{-1}$ throughout. An age of 3.80 ± 0.05 Gyr is assumed for the Isua samples. The effect of the 50-Myr uncertainty on the age is represented by 'error bars' on each data point. An age of 3.70 Gyr is assumed for the Amitsoq gneisses. This figure indicates that the 3.8-Gyr-old mantle had higher Sm/Nd and, as indicated by some samples, also higher Lu/Hf than the modern mantle. The early Archaean mantle is intermediate between the poorly mixed mantles of Mars and the Moon, on the one hand, and the well-mixed mantle of the modern Earth, on the other hand.

Building isotopic differences by radioactive decay would require a time of accretion far longer than is acceptable by current concepts of planetary formation. In addition, abundance variations of the stable isotopes of Hf and Nd are so far undocumented. The BSE values for the Lu–Hf isotope system were determined from a grand average of many chondrites²⁶, whereas those of the Sm–Nd isotope system (chondrite uniform reservoir, CHUR) were based on data only for the Juvinas achondrite²⁵. Because of the excellent agreement between the average $^{147}\text{Sm}/^{144}\text{Nd}$ ratios of chondrites from different classes²⁵ with the CHUR value, there is so far no evidence that the BSE 'cross-hair' shown in Fig. 2 (at 1, 1) and representing the chondritic reference values is significantly misplaced.

We therefore turn to a geodynamic interpretation of the data presented in Fig. 2. The relatively large spread of Nd and Hf isotopic values in the mantle at the time of formation of the Isua rocks may result from a combination of strong elemental fractionation during mantle melting and incomplete mixing by mantle convection. It may also be inherited from the initial stage of accretion combined with initial mantle differentiation. Evidence that the range of Nd isotope heterogeneities was much smaller 2.7 Gyr ago than in modern times—even more than expected from the effect of radiogenic ingrowth—indicates vigorous convection in the late Archaean mantle²⁷. In comparison, the mantles of non-convecting planets, the Moon and Mars, are isotopically extremely heterogeneous. This contrast suggests that the isotopic heterogeneities in the early Archaean terrestrial mantle result from a transient effect of the initial chemical state of the Earth. One metabasalt from Isua was found to contain a measurable anomaly of ^{142}Nd (ref. 8), an isotope produced by the now extinct radionuclide ^{146}Sm which decays with a half-life of 102 Myr. The persistence of such transient chemical heterogeneities for ~0.8 Gyr after planetary accretion is consistent with the present work. These heterogeneities could have been preserved, in spite of a strong convective regime, if mantle layering is stable owing to density stratification and circulation along the surfaces of constant density²⁸.

The nature of the magmatic processes, such as the accumulation of lunar-type cumulates at the bottom of a magma ocean or the formation of a protocrust responsible for the isotopic heterogeneities, has been extensively discussed elsewhere^{9,29}. The time from onset of mantle convection and subsequent thermal relaxation of the Earth until heat production and heat loss balance is effectively achieved is estimated to take several hundreds of Myr (refs 3, 4). Although at least part of the mantle was probably more fractionated 3.8 Gyr ago than today, the geochemical record is far too blurred by later metamorphic disturbances to assess the absolute extent of these mantle heterogeneities. We therefore cannot relate the integrated effect of mantle convection and initial heterogeneity directly to the modern chemical heterogeneity observed in oceanic basalts. The present results, however, confirm that the preservation of particularly strong chemical heterogeneities in the early Archaean mantle reflects the fact that, by that time, the planet had not yet reached thermal steady state.

An important question is whether any primordial differentiated reservoirs have been preserved deep in the modern terrestrial mantle, and whether they could explain the present-day heat flow and rare-gas isotope systematics in basalts. It was recently suggested that an enriched layer (thermal boundary layer) inherited from the primordial mantle differentiation still exists at the base of the mantle²⁸. Direct geophysical evidence for such a layer, which rests essentially on the decreasing correlation length at depths in excess of ~1,600 km and on the splitting of seismic normal modes³⁰, is still relatively faint. Geochemical evidence for a hidden reservoir is indirect, and is based on global terrestrial thermal and chemical inventories with reference to an *a priori* chemical BSE model. The relatively small fraction of surface terrestrial heat flow accounted for by the heat production of radioactive elements in the crust and the upper mantle requires that substantial amounts of U, Th and K be

deeply buried inside the Earth. The subchondritic Nb/La abundances in both the continental crust and the upper mantle³¹, the 'paradox' that the Pb isotopic composition of nearly all terrestrial samples fall to the right of the geochron³² and, more recently²⁶, the Hf and Nd isotopic and elemental imbalance between chondrites, the upper mantle and the continental crust, have all been used to infer a deep missing reservoir. The Archaean meta-igneous rocks from Greenland may therefore carry geochemical evidence for a so far elusive thermal boundary layer before its processing by 3.8 Gyr of mantle convection.

The nature of this reservoir is uncertain²⁸. Neither the survival of a deep layer of primitive material left unperturbed after core segregation nor the segregation of lithospheric plates in the deep mantle⁶, possibly at the core–mantle boundary, could account for the deficit of the heat-producing elements U, Th and K and thereby the 'missing heat'. An alternative model is the segregation of a very dense mineral assemblage rich in Nb, Ta, radioactive, and other elements during an early magma ocean. Although the relative Sm/Nd and Lu/Hf fractionation trends visible in Fig. 2 may help constrain the mineralogy of the segregating phases²⁶, too little is yet known about minor high-pressure minerals that could sequester these elements. □

Received 15 June 1999; accepted 11 February 2000.

1. Staudacher, T. & Allègre, C. J. Terrestrial xenology. *Earth Planet. Sci. Lett.* **60**, 389–406 (1982).
2. Honda, M., McDougall, I., Patterson, D. B., Dougeris, A. & Clague, D. A. Possible solar noble-gas component in Hawaiian basalts. *Nature* **349**, 149–151 (1991).
3. Schubert, G., Stevenson, D. & Cassen, P. Whole planet cooling and the radiogenic heat source contents of the Earth and the Moon. *J. Geophys. Res.* **85**, 2531–2538 (1980).
4. Christensen, U. R. Thermal evolution models for the Earth. *J. Geophys. Res.* **90**, 2995–3007 (1985).
5. Shirey, S. B. & Hanson, G. N. Mantle heterogeneity and crustal recycling in Archaean granite-greenstone belts: evidence from Nd isotopes and trace elements in the Rainy Lake area, Superior province, Ontario, Canada. *Geochim. Cosmochim. Acta* **50**, 2631–2651 (1986).
6. Chase, C. G. & Patchett, P. J. Stored mafic/ultramafic crust and early Archaean mantle depletion. *Earth Planet. Sci. Lett.* **91**, 66–72 (1988).
7. Galer, S. J. G. & Goldstein, S. L. Early mantle differentiation and its thermal consequences. *Geochim. Cosmochim. Acta* **55**, 227–239 (1991).
8. Harper, J., C. L. & Jacobsen, S. B. Evidence from coupled ¹⁴⁷Sm–¹⁴³Nd and ¹⁴⁶Sm–¹⁴²Nd systematics for very early (4.5–Gyr) differentiation of the Earth's mantle. *Nature* **360**, 728–732 (1992).
9. Bennett, V. C., Nutman, A. P. & McCulloch, M. T. Nd isotopic evidence for transient, highly depleted mantle reservoirs in the early history of the Earth. *Earth Planet. Sci. Lett.* **119**, 299–317 (1993).
10. Vervoort, J. D. & Blichert-Toft, J. Evolution of the depleted mantle: Hf isotope evidence from juvenile rocks through time. *Geochim. Cosmochim. Acta* **63**, 533–556 (1999).
11. Rosing, M. T. The theoretical effect of metasomatism on Sm–Nd isotopic systems. *Geochim. Cosmochim. Acta* **54**, 1337–1341 (1990).
12. Gruau, G., Rosing, M., Bridgwater, D. & Gill, R. C. O. Resetting of Sm–Nd systematics during metamorphism of >3.7-Ga rocks: implications for isotopic models of early Earth differentiation. *Chem. Geol.* **133**, 225–240 (1996).
13. Vervoort, J. D., Patchett, P. J., Gehrels, G. E. & Nutman, A. P. Constraints on early Earth differentiation from hafnium and neodymium isotopes. *Nature* **379**, 624–627 (1996).
14. Blichert-Toft, J., Albarède, F., Rosing, M., Frei, R. & Bridgwater, D. The Nd and Hf isotopic evolution of the mantle through the Archaean. Results from the Isua supracrustals, West Greenland, and from the Birimian terranes of West Africa. *Geochim. Cosmochim. Acta* **63**, 3901–3914 (1999).
15. Blichert-Toft, J., Gleason, J. D., Télouk, P. & Albarède, F. The Lu–Hf geochemistry of shergottites and the evolution of the Martian mantle–crust system. *Earth Planet. Sci. Lett.* **173**, 25–39 (1999).
16. Nyquist, L. E. & Shih, C.-Y. The isotopic record of lunar volcanism. *Geochim. Cosmochim. Acta* **56**, 2213–2234 (1992).
17. Unruh, D. M., Stille, P., Patchett, P. J. & Tatsumoto, M. Lu–Hf and Sm–Nd evolution in lunar mare basalts. *J. Geophys. Res.* **89**, B459–B477 (1984).
18. Beard, B. L., Taylor, L. A., Scherer, E. E., Johnson, C. M. & Snyder, G. A. The source region and melting mineralogy of high-titanium and low-titanium lunar basalts deduced from Lu–Hf isotope data. *Geochim. Cosmochim. Acta* **62**, 525–544 (1998).
19. Harper, J., C. L., Nyquist, L. E., Bansal, B., Wiesmann, H. & Shih, C.-Y. Rapid accretion and early differentiation of Mars indicated by ¹⁴²Nd/¹⁴⁴Nd in SNC meteorites. *Science* **267**, 213–217 (1995).
20. Lee, D.-C. & Halliday, A. N. Core formation on Mars and differentiated asteroids. *Nature* **388**, 854–857 (1997).
21. Blichert-Toft, J. & Arndt, N. T. Hf isotope compositions of komatiites. *Earth Planet. Sci. Lett.* **171**, 439–451 (1999).
22. Nutman, A. P., McGregor, V. R., Friend, C. R. L., Bennett, V. C. & Kinny, P. D. The Itsaq Gneiss Complex of southern West Greenland; the world's most extensive record of early crustal evolution (3900–3600 Ma). *Precamb. Res.* **78**, 1–39 (1996).
23. Salters, V. J. M. & White, W. M. Hf isotope constraints on mantle evolution. *Chem. Geol.* **145**, 447–460 (1998).
24. Vervoort, J. D., Patchett, P. J., Blichert-Toft, J. & Albarède, F. Relationships between Lu–Hf and Sm–Nd isotopic systems in the global sedimentary system. *Earth Planet. Sci. Lett.* **168**, 79–99 (1999).
25. Jacobsen, S. B. & Wasserburg, G. J. Sm–Nd isotopic evolution of chondrites. *Earth Planet. Sci. Lett.* **50**, 139–155 (1980).
26. Blichert-Toft, J. & Albarède, F. The Lu–Hf isotope geochemistry of chondrites and the evolution of the mantle–crust system. *Earth Planet. Sci. Lett.* **148**, 243–258 (1997).

27. Blichert-Toft, J. & Albarède, F. Short-lived chemical heterogeneities in the Archaean mantle with implications for mantle convection. *Science* **263**, 1593–1596 (1994).
28. Kellogg, L. H., Hager, B. H. & van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
29. Bowring, S. A. & Housh, T. The Earth's early evolution. *Science* **269**, 1535–1540 (1995).
30. Ishii, M. & Troemp, J. Normal-mode and free-air gravity constraints on lateral variations in velocity and density of the mantle. *Science* **285**, 1231–1236 (1999).
31. McDonough, W. F. Partial melting of subducted oceanic crust and isolation of its residual eclogitic lithology. *Phil. Trans. R. Soc. Lond. A* **335**, 407–418 (1991).
32. Allègre, C. J. Comportement des systèmes U–Th–Pb dans le manteau supérieur et modèle d'évolution de ce dernier au cours des temps géologiques. *Earth Planet. Sci. Lett.* **5**, 261–269 (1969).
33. Dalmasso, J., Barci-Funel, G. & Ardisson, G. J. Reinvestigation of the decay of the long-lived odd-odd ¹⁷⁶Lu nucleus. *Appl. Radiat. Isot.* **43**, 69–76 (1992).
34. Nir-El, Y. & Lavi, N. Measurement of half-life of ¹⁷⁶Lu. *Appl. Radiat. Isot.* **49**, 1653–1655 (1998).

Acknowledgements

We thank S. Goldstein and W. White for comments on the manuscript, and F. Begemann for pointing out the re-determination of the ¹⁷⁶Lu half-life. This work was supported by the Danish Lithosphere Centre, the Carlsberg Foundation, the Danish National Research Fund, the Institut National des Sciences de l'Univers (Programme Cycles Géochimiques), and the Programme National de Planétologie.

Correspondence and requests for materials should be addressed to F.A. (e-mail: albarède@ens-lyon.fr).

Molecular analysis of Neanderthal DNA from the northern Caucasus

Igor V. Ovchinnikov*†‡, Anders Götherström§, Galina P. Romanova||, Vitaliy M. Kharitonov¶, Kerstin Lidén§ & William Goodwin*

* Human Identification Centre, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

† Institute of Gerontology, Moscow 129226, Russia

§ Archaeological Research Laboratory, Stockholm University, 106 91 Stockholm, Sweden

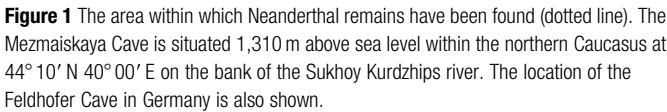
|| Institute of Archaeology, Moscow 117036, Russia

¶ Institute and Museum of Anthropology, Moscow State University, Moscow 103009, Russia

‡ Present address: Department of Medicine, Columbia University, New York, New York 10032 USA

The expansion of premodern humans into western and eastern Europe ~40,000 years before the present led to the eventual replacement of the Neanderthals by modern humans ~28,000 years ago¹. Here we report the second mitochondrial DNA (mtDNA) analysis of a Neanderthal, and the first such analysis on clearly dated Neanderthal remains. The specimen is from one of the eastern-most Neanderthal populations, recovered from Mezmaiskaya Cave in the northern Caucasus². Radiocarbon dating estimated the specimen to be ~29,000 years old and therefore from one of the latest living Neanderthals³. The sequence shows 3.48% divergence from the Feldhofer Neanderthal⁴. Phylogenetic analysis places the two Neanderthals from the Caucasus and western Germany together in a clade that is distinct from modern humans, suggesting that their mtDNA types have not contributed to the modern human mtDNA pool. Comparison with modern populations provides no evidence for the multiregional hypothesis of modern human evolution.

The first successful extraction and sequencing of the mtDNA hypervariable regions (I and II (HVRI & HVRII)) was performed on the Neanderthal-type specimen from Feldhofer Cave, the Neander valley, Germany^{4,5}. Phylogenetic analysis of the sequence placed the Neanderthal mtDNA outside the mtDNA pool of modern humans. This was regarded as a breakthrough in the study of modern human evolution, providing molecular evidence that Neanderthals did not contribute mtDNA to modern humans. From this sequence the divergence of Neanderthals and modern humans was estimated to



have occurred between 317,000 and 741,000 years ago^{4,5}. However, these estimates were based on the molecular analysis of a single specimen. The shortage of potentially well preserved Neanderthal material⁶ and limited access to Neanderthal remains for destructive analysis have hindered the analysis of additional specimens, but genetic characterization of additional Neanderthals is essential to understand their molecular diversity and the relationship between different Neanderthal populations, and to assess their relationship to modern humans further.

The preservation of collagen-type debris was used as an indicator of macromolecule preservation in the bone. The amount of collagen-type debris extracted⁹ from 130 mg of the Mezmaiskaya Neanderthal rib fragment was 22% of the average level extracted from modern bones, and the extracted collagen contained 41.6%

Two sections of one rib (90 mg and 123 mg) were used for DNA extraction in two independent laboratories. In the Glasgow laboratory, a total of 345 base pairs (bp) of HVRI was determined from two overlapping polymerase chain reaction (PCR) fragments with lengths of 232 and 256 bp. Forty PCR amplification cycles produced sufficient product to enable direct sequencing. Products from independent PCR amplifications were also cloned into a TA vector and sequenced (Fig. 2).

Comparison of the 345-bp fragment of HVRI with the Anderson reference sequence¹¹ and the Neanderthal from Feldhofer Cave⁴ revealed 22 differences (17 transitions, 4 transversions and 1 insertion) and 12 differences (11 transitions and 1 transversion), respectively. The Feldhofer Neanderthal HVRI contained 27 differences to Anderson reference sequence¹¹ (over the equivalent 345 bp⁴). The two Neanderthals share 19 substitutions relative to the reference sequence. The cloned PCR products contained all the substitutions that were detected by direct sequencing; six other non-reproducible substitutions occurred in seven different clones. No modern sequences were found in the Glasgow laboratory either by direct sequencing or by sequencing cloned PCR products. The

Figure 2 Variable sites in the DNA sequences of the PCR fragments obtained by direct sequencing (Direct 1 and 2) and cloned PCR products derived from the Neanderthal from Mezmaiskaya Cave. The human reference sequence¹¹ and the sequence of the

NATURE | VOL 404 | 30 MARCH 2000 | www.nature.com

Stockholm laboratory experienced problems with contamination: most of the cloned PCR products that they analysed contained sequences that are found in the modern human mtDNA pool, with two haplotypes predominant. However, three clones contained DNA that was the same as the sequence determined in Glasgow (two of these contained non-reproducible substitutions).

The preservation of 256-bp DNA fragments in bone that is ~29,000 years old, that has not been preserved in permafrost and that contained sufficient DNA to enable direct DNA sequencing after amplification is unprecedented and may be attributed to specific features of the microenvironment of the limestone cavern². The retrieval of mtDNA showed a positive correlation to the preservation of collagen content and the skeletal morphology.

Phylogenetic analysis using both distance and parsimony optimizations places the two Neanderthal sequences together, in a distinct clade, basal to modern humans. Neighbour-joining analysis supports this separation (Fig. 3a). Parsimony analysis, which makes minimal assumptions about the model of evolution and optimizes the fit between the tree and data, produced similar results (Fig. 3b).

The level of pairwise difference found between the two Neanderthals was higher than the average values found in random samples of 300 Caucasoids (5.28 ± 2.24) and Mongoloids (6.27 ± 2.29)—less than 1% of Caucasoid and Mongoloid pairs differ at 12 or more positions—but comparable to a random sample

of 300 Africans (8.36 ± 3.2), where 37% of pairs differed at 12 or more positions. When analysing ancient DNA there is the possibility of misincorporating nucleotides in the early stages of PCR, especially when the target DNA is possibly damaged and present in low copy number¹². As both Neanderthals were analysed in replicate and the results were consistent, however, errors of this type can be discounted.

The Feldhofer and Mezmaiskaya Neanderthals were separated geographically by over 2,500 km. Given that these two individuals contained closely related mtDNA, which is phylogenetically distinct from modern humans, and displays only a moderate level of sequence diversity compared with some primates¹³, these data provide further support for the hypothesis of a very low gene flow between the Neanderthals and modern humans. In particular, these data reduce the likelihood that Neanderthals contained enough mtDNA sequence diversity to encompass modern human diversity.

The 'out-of-Africa' hypothesis for the origin of modern humans predicts equal distances between the Neanderthal sequences and all modern sequences. We observed this in our analysis—the average pairwise differences between the Neanderthals and 300 randomly selected Africans, Mongoloids and Caucasoids were calculated to be 23.09 ± 2.86 , 23.27 ± 4.06 and 25.45 ± 3.27 , respectively.

We estimated the age of the most recent common ancestor (MRCA) of the mtDNA of the eastern and western Neanderthals to be 151,000–352,000 years. This coincides with the time of emergence of the Neanderthal lineage in the palaeontological records¹⁴. The divergence of modern human and Neanderthal mtDNA was estimated to be between 365,000 and 853,000 years. Using the same model, we estimated the age of the earliest modern human divergences in mtDNA to be between 106,000 and 246,000 B.P.

The results obtained from this specimen suggest that some other Neanderthal samples may be amenable to molecular analysis. To obtain a more complete picture of the relationship of Neanderthals to modern humans, additional Neanderthals and early modern humans must be analysed, especially from the regions where they may have co-existed. The excellent preservation of this specimen leads to the potential of analysing the entire Neanderthal mitochondrial genome. □

Methods

DNA extraction, PCR, cloning and sequencing

The DNA extraction methods used in Glasgow⁴ and Stockholm¹⁵ have been described. We took precautions to prevent contamination from modern DNA⁴. The Neanderthal-specific primer pairs, NL16,055 and NH16,262, and NL16,209 and NH16,400 (5'-TGATTTCAC GGAGGATGGTGA-3') were used in Glasgow and the primers L16,212 (5'-ATGCTTAC AAGCAAGCACA-3') and H16,332 (5'-TTGACTGTAATGTGCTATG-3') were used in Stockholm. The annealing temperatures for the primer pairs NL16,055–NH16,262, NL16,209–NH16,400 and L16,212–H16,332 were 50 °C, 60 °C and 50 °C respectively; 40 cycles were used for the first two pairs, 55 cycles for the third. AmpliTaq Gold (Perkin Elmer Cetus) was used in all PCRs. PCR products were purified using the QIAquick Gel Extraction kit (Qiagen) before direct sequencing using the Dye Terminator sequencing kit (Perkin Elmer) or cloning into the TA vector (Invitrogen) before sequencing with the same kit using the M13 and T7 primers.

Sequence analyses

The neighbour-joining and the maximum parsimony branch and bound trees were both constructed using PAUP* 4.0 (ref. 16). For the neighbour-joining analysis, the Tamura-Nei DNA substitution model¹⁷ was used with a gamma distribution of 0.4 (ref. 18), for all other parameters the defaults provided by PAUP* 4.0 were used. The MRCA was calculated using the described methods and assumptions⁵. PAUP* 4.0 was used to calculate pairwise differences between sequences: the data sets used for this were constructed by randomly selecting appropriate samples from a published data set¹⁹.

Received 15 November 1999; accepted 31 January 2000.

1. Stringer, C. B. & Mackie, R. *African Exodus: the Origin of Modern Humanity* (Cape, London, 1996).
2. Golovanova, L. V., Hofferer, J. F., Kharitonov, V. M. & Romanova, G. P. Mezmaiskaya Cave: A Neanderthal occupation in the Northern Caucasus. *Curr. Anthropol.* **40**, 77–86 (1999).
3. Smith, F. H., Trinkaus, E., Pettitt, P. B., Karavanic, I. & Paunovic, M. Direct radiocarbon dates for Vindija G1 and Velika Pecina Late Pleistocene hominid remains. *Proc. Natl Acad. Sci. USA* **96**, 12281–12286 (1999).

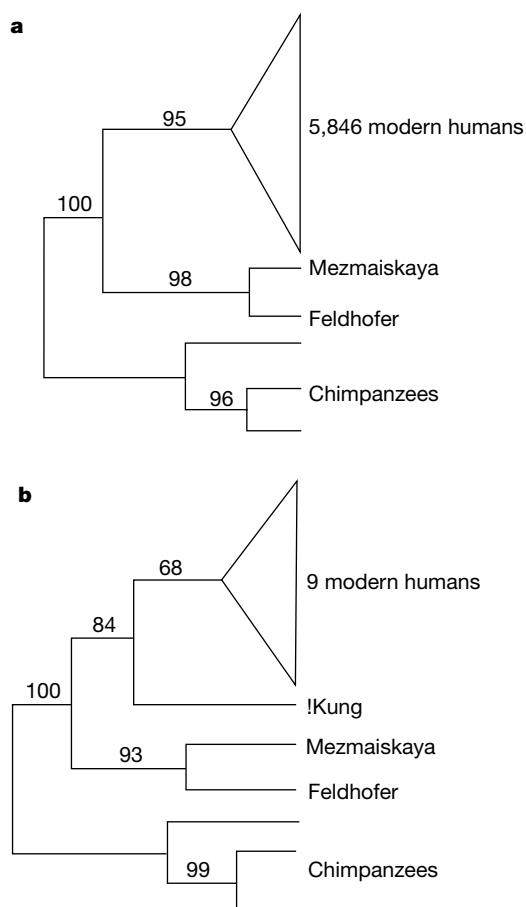


Figure 3 Phylogenetic relationship of the two Neanderthals and modern humans. **a**, A neighbour-joining tree computed using a total of 1,897 haplotypes derived from 5,846 modern humans¹⁹. **b**, A maximum parsimony branch and bound search with the two Neanderthal sequences along with the sequences of one !Kung, three other Africans, three Asians and three Europeans, all randomly selected¹⁹. This result is congruent with four additional data sets that were analysed. In both analyses, three chimpanzee sequences¹³ were used as an outgroup. The numbers in both diagrams refer to the bootstrap frequencies (%) obtained from 1,000 replicates.

4. Krings, M. *et al.* Neandertal DNA sequence and the origin of modern humans. *Cell* **90**, 19–30 (1997).
5. Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H. & Pääbo, S. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. *Proc. Natl Acad. Sci. USA* **96**, 5581–5585 (1999).
6. Cooper, A. *et al.* Neandertal genetics. *Science* **277**, 1021–1023 (1997).
7. Gabunia, L. & Vekua, A. A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus. *Nature* **373**, 509–512 (1995).
8. Kozłowski, J. K. in *Neandertals and Modern Humans in Western Asia* 461–482 (Plenum, New York–London, 1998).
9. Brown, T. A., Nelson, D. E., Vogel, J. S. & Southon, J. R. Improved collagen extraction by modified Longin method. *Radiocarbon* **30**, 171–177 (1988).
10. DeNiro, M. J. Postmortem preservation and alteration of *in vivo* bone collagen isotope ratios in relation to palaeodietary reconstruction. *Nature* **317**, 806–809 (1985).
11. Anderson, S. *et al.* Sequence and organisation of the human mitochondrial genome. *Nature* **290**, 457–474 (1981).
12. Hoss, M. *et al.* DNA damage and DNA sequence retrieval from ancient tissue. *Nucleic Acids Res.* **24**, 1304–1307 (1996).
13. Gagneux, P. *et al.* Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc. Natl Acad. Sci. USA* **96**, 5077–5082 (1999).
14. Gamble, C. in *Prehistoric Europe* 5–41 (Oxford Univ. Press, Oxford, 1998).
15. Lidén, K., Götherström, A. & Eriksson, E. Diet, gender and rank. *ISKOS* **11**, 158–164 (1997).
16. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)* Version 4. (Sinauer Associates, Sunderland, Massachusetts, 1998).
17. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *J. Mol. Evol.* **10**, 512–526 (1993).
18. Excoffier, L. & Yang, Z. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol. Biol. Evol.* **16**, 1357–1368 (1999).
19. Burckhardt, F., von Haeseler, A. & Meyer, S. HvrBase: compilation of mtDNA control region sequences from primates. *Nucleic Acids Res.* **27**, 138–142 (1999).

Acknowledgements

We are indebted to L. V. Golovanova for the excavations in Mezmaiskaya Cave that provided materials for analysis. We thank V. P. Ljubin and P. Vanezis for encouragement and support; B. L. Cohen for numerous discussions; J. L. Harley, O. I. Ovtchinnikova, E. B. Druzina and J. Wakefield for technical help and assistance; R. Page for help with the phylogenetic analysis; and P. Beerli, A. Cooper, M. Cusack, M. Nordborg and M. Ruvolo for useful comments. I.V.O. thanks his host G. Curry. I.V.O. was supported by a Royal Society/NATO Fellowship. We thank the Swedish Royal Academy of Sciences and the Swedish Research Council for Natural Sciences for partial financial support.

Correspondence and requests for material should be addressed to W.G.
(e-mail: w.goodwin@formed.gla.ac.uk).

Pervasive density-dependent recruitment enhances seedling diversity in a tropical forest

Kyle E. Harms, S. Joseph Wright, Osvaldo Calderón, Andrés Hernández & Edward Allen Herre

Smithsonian Tropical Research Institute, Apartado 2072, Balboa, Republic of Panamá

Negative density-dependent recruitment of seedlings, that is, seeds of a given species are less likely to become established seedlings if the density of that species is high, has been proposed to be an important mechanism contributing to the extraordinary diversity of tropical tree communities^{1–3} because it can potentially prevent any particular species from usurping all available space, either in close proximity to seed sources or at relatively larger spatial scales^{1–18}. However, density-dependent recruitment does not necessarily enhance community diversity¹⁴. Furthermore, although density-dependent effects have been found at some life stages in some species^{3–13}, no study has shown that density-dependent recruitment affects community diversity^{14,15}. Here we report the results of observations in a lowland, moist forest in the Republic of Panamá in which the species identities of 386,027

seeds that arrived at 200 seed traps were compared with the species identities of 13,068 seedlings that recruited into adjacent plots over a 4-year period. Across the 200 sites, recruit seedling diversity was significantly higher than seed diversity. Part of this difference was explained by interspecies differences in average recruitment success. Even after accounting for these differences, however, negative density-dependent recruitment contributes significantly to the increase in diversity from seeds to seedling recruits.

Here, both within and between species, we link seed input to initial seedling recruitment in the diverse tropical forest on Barro Colorado Island, Panamá (see Methods). With these data, we first estimate the strength of density-dependent recruitment for individual species and show that it is both pervasive and strong across the community. Then, to determine the influence of these effects on diversity, we compare the observed diversity of seedling recruits to (1) the observed diversity of seeds, (2) the expected diversity of seedling recruits after accounting for average differences in recruitment success among species, but in the absence of density dependence, and (3) the expected diversity of seedling recruits after incorporating our species-specific estimates of density dependence. We thereby test the central and previously unexamined issue of whether density-dependent recruitment increases species diversity.

Seedling recruitment was strongly, negatively density-dependent in our 53 focal species (data for *Trichilia tuberculata*, the most common canopy tree at the site, are shown in Fig. 1). Slopes for the seed–recruit regressions were less than 1 for every species, with a median value of just 0.23 (Fig. 2). Nevertheless, high initial seed densities often overwhelmed density-dependent differences in per seed recruitment so that recruit density was greatest where seeds were initially most numerous; regression slopes were weakly positive for 41 of the 53 species (Fig. 2). To determine the net effect on diversity of these potentially conflicting influences, we therefore conducted the following comparisons.

We first compared the observed diversity of seeds with the observed diversity of seedling recruits at each station. Considering only the 53 focal species, the average Shannon–Wiener diversity index was significantly lower for seeds than for recruits (1.33 and 1.95, respectively, paired *t*-test = 14.7, *P* < 0.0001; Fig. 3a). In addition, the local diversities of seeds and recruits were unrelated

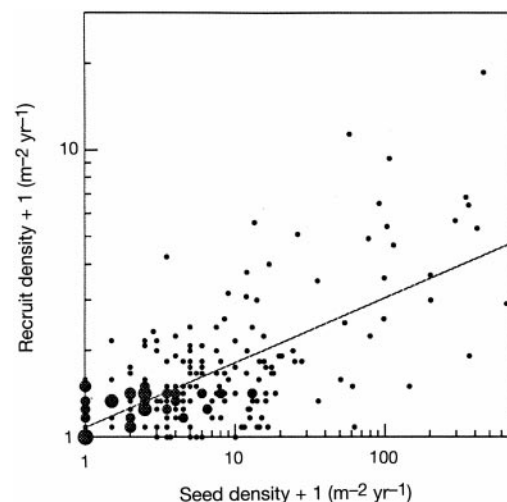


Figure 1 The relationship between the seed density and recruit seedling density for *Trichilia tuberculata* (Meliaceae). The slope of the log–log relationship is less than 1, indicating that recruitment is negatively density dependent. Nonetheless, recruit density increases with seed density. Each symbol represents a census station(s) consisting of one 0.5-m² seed trap and three 1-m² seedling plots. Symbol size is proportional to the number of stations (ranging from one to six) with identical counts of seeds and recruits.

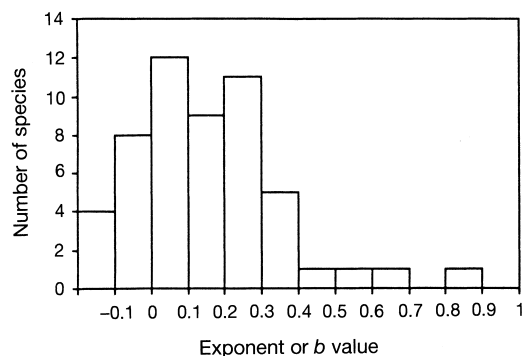


Figure 2 The frequency distribution of the exponent of the relationship between recruit density and seed density for 53 species of shrubs, trees and lianas. All exponents were less than 1 and the median was 0.23, suggesting that strong negative density dependence characterizes seedling recruitment.

(correlation coefficient, $r^2 = 0.0029$); therefore, for these well-sampled species, diversity increased markedly during the seed-to-seedling transition, and the diversity of the local seed rain was a very poor indicator of local seedling diversity. Similarly, the Shannon–Wiener index was lower for the aggregate of all seeds from all stations (2.82) as compared with the aggregate of all recruits (3.80). Moreover, the observed recruit diversity at the 200 stations was very similar to the overall diversity of the 229,071 stems greater than 1-cm diameter at breast height in the 50-Ha plot (Shannon–Wiener index = 3.93)¹⁹.

We then explored the relationship between recruit diversity and seed diversity using species specific, seed–recruit regressions to predict community-level seedling diversity. However, the species-specific regressions incorporate both density dependence and mean seed-to-seedling transition probabilities. In tropical forests, fecund species tend to produce small seeds that require relatively high light intensities to establish successfully, whereas less fecund species tend to produce larger seeds that establish readily in the shaded understory^{20,21}. This life-history variation may increase diversity during the seed-to-seedling transition as small-seeded species tend to produce high seed densities and have low seed-to-seedling transition probabilities in the understory. As an example, five species with a seed mass of less than 8 mg accounted for 65.5% of the seed rain but just 5.0% of the recruits in this study. To isolate the consequences of these interspecies differences in average recruitment success, we calculated an expected recruit species composition for each station by multiplying the number of seeds of each species by a species-specific, mean seed-to-seedling transition probability (estimated as mean recruit density divided by mean seed density for each species). The expected diversity of recruits (mean = 1.65) was greater than the observed diversity of seeds (paired t -test = 6.92, $P < 0.0001$) but less than the observed diversity of recruits (paired t -test = 8.32, $P < 0.0001$; Fig. 3b). Therefore, differences among species in their average success of seedling establishment accounted for part, but not all, of the difference in diversity observed between the seed rain and seedling recruits.

To estimate the influence of species-specific density dependence, we calculated a second expected recruit species composition for each station by substituting the number of seeds of each species into the appropriate, species-specific seed–recruit regression. The expected station diversities incorporating these density-dependent transition probabilities (mean = 1.91) were not significantly different from the observed recruit diversities (paired t -test = 1.63, $P > 0.10$; Fig. 3c). Interspecies differences in average recruitment success explained part of the increase in diversity observed during the seed-to-seedling transition. Strong and pervasive density dependence explained the remainder.

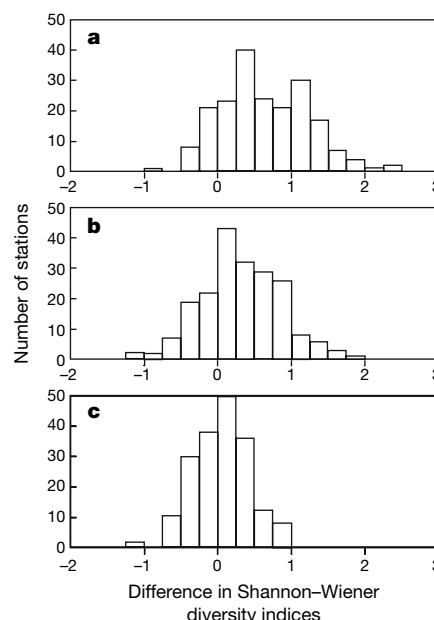


Figure 3 Comparisons of Shannon–Wiener diversity indices within census stations. The diversity of recruits minus the diversity of the seed rain (a), the diversity of recruits predicted by incorporating interspecific variation in average recruitment success (b), and the diversity of recruits predicted by incorporating both interspecific differences in average recruitment success and density dependence (c). Predicted recruit diversity converges on observed recruit diversity with the inclusion of species-specific differences in average recruitment success (b) and density dependence (c).

Several studies have shown high levels of host specificity in herbivorous insects and other pests that consume either leaf or seed tissue of tropical plants; such pests could be involved in producing the observed density-dependent effects^{4,7,10,22,23}. Together with our findings, these studies suggest that the high diversity of plant species in this tropical forest is at least partially due to the continuing depredations by the pests and pathogens that selectively kill seeds and seedlings. Moreover, greater pest and pathogen pressure in the tropics^{17,24,25}, together with the potential effect on plant regeneration shown here, may contribute to the well-known latitudinal gradient in plant diversity. □

Methods

Establishing and monitoring census stations

In December 1986, 200 seed traps were erected within the Forest Dynamics Project plot of Barro Colorado Island, Panama^{26,27}. The average distance ($\pm$ s.d.) between seed traps was 18.9 ± 3.6 m. Each trap consisted of a 0.5-m², open-topped, 1-mm wire-mesh bag held 0.8 m above the ground on a PVC frame. All seeds falling into the traps were counted and identified to species each week from 1 January 1987 to 31 December 1998 (ref. 27). Between January and March 1994, all woody plants less than 50 cm tall were tagged and identified to species in 600 1-m² plots. These plots were placed 2 m from each of three sides of each seed trap. Survivors were re-measured and new recruits were tagged and identified between January and March each year from 1995 to 1998. Each seed trap and its three adjacent plots constitute a census station.

Over 4 years, the seed rain included 287 species that germinate in soil and averaged 965 seeds m⁻² yr⁻¹. During the study, most of the 200 census stations never received seeds from most species, in part due to low densities of reproductive adults, variable fecundity and limited seed dispersal^{17,27,28}. First-year recruits included 216 species and averaged 5.45 individuals m⁻² yr⁻¹. Seedling turnover was high, with an average of 55% of a year's recruits dying in their first year.

Regression analyses

To estimate the strength of density dependence on seedling recruitment, we examined the relationship ($R = aS^b$) between the density of recruits (R) and the density of seeds (S). Data were pooled across years within census stations, and care was taken to associate the appropriate years of seed fall and recruitment, especially for species with delayed

germination. We evaluated the seed–recruit relationship using linear regression of log transformed values of $R + 1$ and $S + 1$ to normalize residuals. In these regressions, the fitted constant b takes values of less than 1 if the per seed probability of recruitment is inversely related to seed density.

Species were excluded from regression analyses if (1) seeds passed through the 1-mm trap mesh, (2) seeds or recruits were recorded at fewer than 10 stations, or (3) seed or recruit density varied less than a factor of four among stations. Fifty-three species remained including 28 trees, 20 lianas and five shrubs. These 53 focal species account for 60.3% of the seed rain and 80.4% of recruits. None of the 53 species persist for more than one year in the soil seed bank.

To insure that the regression results were not spurious or artefactual, we conducted several additional tests. First, some traps received many more seeds than could possibly recruit into adjacent seedling plots (simple space limitation). Second, in many instances, seedlings of a given species recruited into some seedling plots, but no seeds of that species were found in the adjacent seed traps. Either or both of these effects would lower b -value estimates and inflate the apparent importance of density dependence. Third, empty stations, lacking both seeds and recruits, introduce an opposing bias, and raise the b -value estimates. To control for the effect of space limitation, we conducted analyses excluding stations in which single species seed densities exceeded the maximum single species recruit density ($66 \text{ seedlings m}^{-2}$). To control for the other biases, we conducted analyses after excluding stations lacking conspecific seed. We then eliminated all of these effects by including only stations with non-zero seed densities less than 66 seed m^{-2} . Finally, the addition of 1 before taking logarithms gives b values a downward bias. This effect can be large, particularly when the ratio of seeds to recruits is large. To evaluate the effect of this final potential bias, we compared 95% confidence intervals of the observed b values with b values predicted in the absence of density dependence. Observed b values were significantly less than predicted b values for every species. In all cases, these more conservative tests substantiated the pervasive and strong effects of negative density dependence on seedling recruitment.

Received 21 December 1999; accepted 10 February 2000.

1. Ridley, H. N. *The Dispersal of Plants Throughout the World* (L. Reeve & Co., Ashford, England, 1930).
2. Janzen, D. H. Herbivores and the number of tree species in tropical forests. *Am. Nat.* **104**, 501–528 (1970).
3. Connell, J. H. in *Dynamics of Populations* (eds den Boer, P. J. & Gradwell, G. R.) 298–312 (Center for Agricultural Publication and Documentation, Wageningen, The Netherlands, 1971).
4. Augspurger, C. K. Seedling survival of tropical tree species: interactions of dispersal distance, light-gaps and pathogens. *Ecology* **65**, 1705–1712 (1984).
5. Clark, D. A. & Clark, D. B. Spacing dynamics of a tropical rainforest tree: evaluation of the Janzen–Connell model. *Am. Nat.* **124**, 769–788 (1984).
6. Hammond, D. S. & Brown, V. K. in *Dynamics of Tropical Communities* (eds Newbery, D. M., Prins, H. H. T. & Brown, N. D.) 51–78 (Blackwell Science, London, 1998).
7. Howe, H. F. Survival and growth of juvenile *Virola surinamensis* in Panama: effects of herbivory and canopy closure. *J. Trop. Ecol.* **6**, 259–280 (1990).
8. Webb, C. O. & Peart, D. R. Seedling density dependence promotes coexistence of Bornean rain forest trees. *Ecology* **80**, 2006–2017 (1999).
9. Condit, R., Hubbell, S. P. & Foster, R. B. Recruitment near conspecific adults and the maintenance of tree and shrub diversity in a neotropical forest. *Am. Nat.* **140**, 261–286 (1992).
10. Gilbert, G. S., Hubbell, S. P. & Foster, R. B. Density and distance-to-adult effects of a canker disease of trees in a moist tropical forest. *Oecologia* **98**, 100–108 (1994).
11. Hubbell, S. P., Condit, R. & Foster, R. B. Presence and absence of density dependence in a neotropical tree community. *Phil. Trans. R. Soc. Lond. B* **330**, 269–281 (1990).
12. Wills, C. & Condit, R. Similar non-random processes maintain diversity in two tropical rainforests. *Proc. R. Soc. Lond. B* **266**, 1445–1452 (1999).
13. Wills, C., Condit, R., Foster, R. B. & Hubbell, S. P. Strong density- and diversity-related effects help to maintain tree species diversity in a neotropical forest. *Proc. Natl Acad. Sci. USA* **94**, 1252–1257 (1997).
14. Hubbell, S. P. Seed predation and the coexistence of tree species in tropical forests. *Oikos* **35**, 214–229 (1980).
15. Wright, S. J. in *Handbook of Functional Plant Ecology* (eds Pugnaire, F. I. & Valladares, F.) 449–472 (M. Dekker, New York, 1999).
16. Connell, J. H. Diversity in tropical rain forest and coral reefs. *Science* **199**, 1302–1309 (1978).
17. Leigh, E. G. Jr *Tropical Forest Ecology* (Oxford Univ. Press, Oxford, 1999).
18. Schupp, E. W. The Janzen–Connell model for tropical tree diversity: population implications and the importance of spatial scale. *Am. Nat.* **140**, 526–530 (1992).
19. Condit, R., Hubbell, S. P. & Foster, R. B. Changes in tree species abundance in a Neotropical forest: impact of climate change. *J. Trop. Ecol.* **12**, 231–256 (1996).
20. Hammond, D. S. & Brown, V. K. Seed size of woody plants in relation to disturbance, dispersal, soil type in wet neotropical forests. *Ecology* **76**, 2544–2561 (1995).
21. Kitajima, K. Relative importance of photosynthetic traits and allocation patterns as correlates of seedling shade tolerance of 13 tropical trees. *Oecologia* **98**, 419–428 (1994).
22. Janzen, D. H. Specificity of seed-attacking beetles in a Costa Rican deciduous forest. *J. Ecol.* **68**, 929–952 (1980).
23. Barone, J. A. Host-specificity of folivorous insects in a moist tropical forest. *J. Anim. Ecol.* **67**, 400–409 (1998).
24. Coley, P. D. & Barone, J. A. Herbivory and plant defenses in tropical forests. *Annu. Rev. Ecol. Syst.* **27**, 305–335 (1996).
25. Coley, P. D. & Aide, T. M. in *Plant–Animal Interactions* (eds Price, P. W., Lewinsohn, T. M., Fernandes, G. W. & Benson, W. W.) 25–49 (John Wiley and Sons, New York, 1991).
26. Hubbell, S. P. *et al.* Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science* **283**, 554–557 (1999).
27. Wright, S. J., Carrasco, C., Calderón, O. & Paton, S. The El Niño southern oscillation, variable fruit production, and famine in a tropical forest. *Ecology* **80**, 1632–1647 (1999).
28. Wright, S. J., Carrasco, C., Calderón, O. & Paton, S. The El Niño southern oscillation, variable fruit production, and famine in a tropical forest. *Ecology* **80**, 1632–1647 (1999).

Acknowledgements

We dedicate this study to the memory of Eduardo Sierra: his ability to identify seedlings of 700 plant species was indispensable. We thank B. Arnold, J. Connell, J. Dalling, J. Eberhard, C. Gehring, P. Green, S. Hubbell, P. Juniper, H. Müller-Landau and T. Theimer for constructive criticism of the manuscript. The Andrew W. Mellon Foundation and the Environmental Sciences Program of the Smithsonian Institution provided financial support.

Correspondence and requests for materials should be addressed to S.J.W. (e-mail: wrightj@tivoli.si.edu).

The evolution of syntactic communication

Martin A. Nowak*, Joshua B. Plotkin* & Vincent A. A. Jansen†

* Institute for Advanced Study, Princeton, New Jersey 08540, USA

† School of Biological Sciences, Royal Holloway, University of London, Egham Surrey, TW20 0EX UK

Animal communication is typically non-syntactic, which means that signals refer to whole situations^{1–7}. Human language is syntactic, and signals consist of discrete components that have their own meaning⁸. Syntax is a prerequisite for taking advantage of combinatorics, that is, “making infinite use of finite means”^{9–11}. The vast expressive power of human language would be impossible without syntax, and the transition from non-syntactic to syntactic communication was an essential step in the evolution of human language^{12–16}. We aim to understand the evolutionary dynamics of this transition and to analyse how natural selection can guide it. Here we present a model for the population dynamics of language evolution, define the basic reproductive ratio of words and calculate the maximum size of a lexicon. Syntax allows larger repertoires and the possibility to formulate messages that have not been learned beforehand. Nevertheless, according to our model natural selection can only favour the emergence of syntax if the number of required signals exceeds a threshold value. This result might explain why only humans evolved syntactic communication and hence complex language.

The uniqueness of language has been compared to that of the elephant’s trunk¹³. Human language is as different from animal communication as the elephant’s trunk is from other animals’ nostrils. Yet few biologists worry about the evolution of the elephant’s trunk (which is a most complex organ that consists of about 6,000 individual muscles and that can perform an unparalleled variety of mechanical tasks), whereas many philosophers, linguists and biologists have great difficulties in imagining how language could have arisen by darwinian evolution^{17–21}.

A challenge for evolutionary biology, therefore, is to provide a detailed mathematical account of how natural selection can enable the emergence of human language from animal communication. Animal communication is based on three basic designs: a finite repertoire of calls (territorial calls or warning of predators); a continuous analogue signal (for example, the dance of bees); and a series of random variations on a theme (such as the song of birds). All natural animal communication appears to be non-syntactic; some caution, however, seems appropriate as the final verdict on complex vocalization patterns of certain primate species or dolphins has not been reached. In contrast, human language is clearly syntactic: messages consist of components that have their own meaning. We compare non-syntactic and syntactic communication and evaluate their relative performance in an evolutionary setting.

First, we formulate a mathematical model for the population dynamics of language evolution. Suppose a language contains n

words. Each individual is born not knowing any of the words, but can acquire words by learning from other individuals. Individuals are characterized by the subset of words that they know. The general equations for the resulting evolutionary dynamics are complicated (see Methods), but an analytical approach is possible if we describe the process in terms of independent, elementary steps on the basis of two assumptions: first, in any one interaction between two individuals only a single new word can be learned; second, words are memorized independently of each other. With these assumptions, we obtain for the population dynamics of x_i , which is the relative abundance of individuals who know word W_i

$$\dot{x}_i = R_i x_i (1 - x_i) - x_i \quad (1)$$

where $i = 1, \dots, n$. The abundance of word W_i spreads by the interaction of people who know the word with people who do not know the word; hence its rate of increase is proportional to the product $x_i(1 - x_i)$. The rate constant, $R_i = bq\phi_i$, is the basic reproductive ratio of word W_i . This is the average number of individuals who acquire word W_i from one individual who knows it. The parameter b is the total number of word-learning events per individual per lifetime. The parameter q is the probability of memorizing a single word after one encounter, and ϕ_i is the frequency of occurrence of word W_i in the (spoken) language. The term $-x_i$ denotes a constant death rate, setting the average lifetime of each individual as one time unit.

For a word to be maintained in the lexicon of the population, its basic reproductive ratio must exceed one, which implies that $\phi_i > 1/(bq)$. Suppose W_i is the least frequent word. We certainly know that ϕ_i is less than $1/n$, which is the frequency of a word if all words have the same frequency. Thus the maximum number of words is $n_{\max} = bq$. Note that this number is always less than the total number of words, b , that are presented to a learning individual. Hence, the lexicon of the population cannot exceed the total number of word-learning events for each individual.

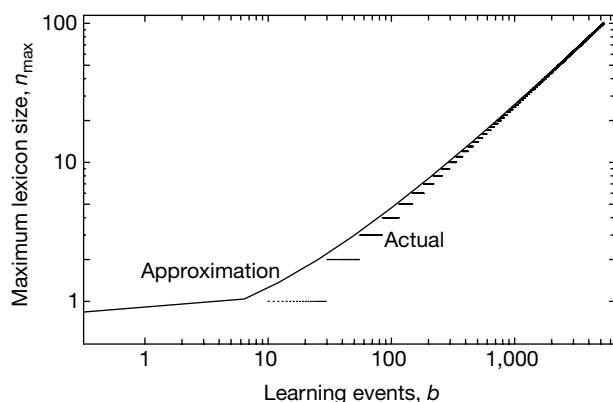


Figure 1 How many word-learning events per individual are required for a population to maintain a certain number of words in its combined lexicon assuming that word frequencies follow Zipf's law? Zipf's law states that the frequency ϕ_i of occurrence of a word is inversely proportional to its position, i , in a frequency ranking. Zipf's law seems to be a very good approximation for every human language. Nobody knows its significance. It can be shown, however, that a random source emitting symbols and spaces also generates word frequency distributions following Zipf's law²³. Assuming that n words follow a frequency distribution as in Zipf's law, $\phi_i = C_n/i$ where $C_n = 1/\sum_{i=1}^n (1/i)$, the maximum number of words that can be maintained, $n_{\max}$, is the largest integer, n , which fulfils the inequality $\phi_n = C_n/n > 1/bq$. The figure shows $n_{\max}$ versus b for $q = 0.1$. The exact relationship is a step function because both b and $n_{\max}$ are integers. The equation $n_{\max}(\gamma + \ln n_{\max}) = bq$ is an excellent approximation (continuous line). Here we have used Euler's gamma which is $\gamma = 0.5772\dots$. For example, to maintain a lexicon of 100 words, we need about $bq = 520$ successful word-learning events per individual. For $q = 0.1$, this implies that a total of roughly $b = 5,200$ word-learning events per individual are required.

Assuming that word frequency distributions follow Zipf's law^{22,23}, $\phi_i = C/i$, where C is a constant, we find that the maximum number of words is roughly given by the equation $n_{\max} \ln(n_{\max}) = bq$ (Fig. 1).

We now use this mathematical framework to analyse how natural selection can guide the transition from non-syntactic to syntactic communication. Imagine a group of individuals who communicate about events in the world around them. Events are combinations of objects, places, times and actions. (We use 'object' and 'action' in a general way to represent everything that can be referred to by nouns and verbs of current human languages.) For notational simplicity, suppose that each event consists of one object and one action. Thus, event E_{ij} consists of object i and action j . Non-syntactic communication uses words for events, whereas syntactic communication uses words for objects and actions (Fig. 2). Events occur at different rates, which are specified by the entries of an 'event rate matrix', Γ .

For natural selection to operate on language design, language must confer fitness. A plausible assumption is that correct communication about events provides a fitness advantage to the interacting individuals. In terms of our model, the fitness contribution of a language can be formulated as the probability that two individuals know the correct word for a given event summed over all events and weighted with the rate of occurrence of these events.

The population dynamics of non-syntactic communication are again given by equation (1) with word W_{ij} referring to event E_{ij} . As before, the maximum number of words that can be maintained in the population is limited by bq . We calculate the fitness of individuals using non-syntactic communication (see Methods).

We now turn to syntactic communication. Noun N_i refers to object i and verb V_j refers to action j , hence the event E_{ij} is described by the sentence $N_i V_j$. For the basic reproductive ratios we obtain $R(N_i) = (b/2)q_s \phi(N_i)$ and $R(V_j) = (b/2)q_s \phi(V_j)$. Here $\phi(N_i)$ and $\phi(V_j)$ denote the frequency of occurrence of noun N_i and verb V_j , respectively. The factor $1/2$ appears because either the noun or the verb is learned in any one of the b learning events. The probability of memorizing a noun or a verb is given by q_s . We expect q_s to be smaller than q , which simply means that it is a more difficult task to learn a syntactic signal than a non-syntactic signal. For both signals, the (arbitrary) meaning has to be memorized; for a syntactic signal one must also memorize its relation to other signals (whether it is a noun or a verb, for example).

For noun N_i to be maintained in the lexicon of the population, its basic reproductive ratio must exceed one, implying that $\phi(N_i) > 2/(bq_s)$. Similarly, for verb V_j we find $\phi(V_j) > 2/(bq_s)$. This means that the total number of nouns plus verbs is limited by bq_s , which is always less than b . The maximum number of grammatical sentences, however, which consist of one noun and one verb, is given by $(bq_s)^2/4$. Hence syntax makes it possible to maintain more sentences than the total number of sentences, b , encountered by a learning individual. All words have to be learned, therefore, but syntactic signals enable the formulation of 'new' sentences that have not been learned beforehand.

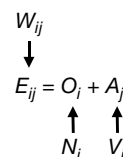


Figure 2 To understand the essence of the evolution of syntax, we imagine a world where each event consists of one object and one action. Event E_{ij} consists of object O_i and action A_j . A non-syntactic language has words, W_{ij} , that refer to events, E_{ij} . A syntactic language has words for objects and actions, N_i and V_j . Words for objects are called nouns, words for actions are called verbs. Our mathematical analysis can also be adapted to more complicated situations, where events consist of several objects, actions, places or times, but the equations become more clumsy. The principles remain the same.

For calculating the fitness of syntactic communication, note that two randomly chosen individuals can communicate about event E_{ij} if they both know noun N_i and verb V_j . If we denote the relative abundance of individuals who know N_i and V_j by $x(N_i V_j)$, we obtain

$$\begin{aligned} \dot{x}(N_i V_j) = & R(N_i)x(N_i)[x(V_j) - x(N_i V_j)] \\ & + R(V_j)x(V_j)[x(N_i) - x(N_i V_j)] - x(N_i V_j) \end{aligned} \quad (2)$$

The abundance of individuals who know noun N_i and verb V_j increases if someone who knows N_i meets someone who knows V_j but not $N_i V_j$. Similarly, the abundance increases if someone

who knows verb V_j meets someone who knows N_i but not $N_i V_j$. We calculate the equilibrium abundances and thence the fitness of individuals using syntactic communication (see Methods). Figure 3 shows the fitness of syntactic and non-syntactic communication as a function of b for different examples of the event rate matrix, Γ .

When does syntactic communication lead to a higher fitness than non-syntactic communication? Suppose there are n objects and m actions. Suppose a fraction, p , of these mn events occur (all at the same frequency), and the other events do not occur. In this case, $R(W_{ij}) = bq/(pmn)$ for those events that occur. Making the (somewhat rough) assumption that all nouns and all verbs, respectively, occur on average at the same frequency, we obtain $R(N_i) = bq_s/(2n)$ and $R(V_j) = bq_s/(2m)$. If all involved basic reproductive ratios are well above one, the fitness of syntactic communication exceeds the fitness of non-syntactic communication provided $(m^2n + mn^2)/(m^2 + mn + n^2) > (2q)/(pq_s)$ (see Methods). If this inequality holds then syntactic communication will be favoured by natural selection; otherwise non-syntactic communication will win. Observe that $m \geq 3$ and $n \geq 3$ are necessary conditions for the evolution of syntax. For $m = n$, the relevant condition is

$$n > 3q/(pq_s) \quad (3)$$

Hence, the size of the system has to exceed a critical value for syntactic communication to evolve. For example, if it is twice as hard to memorize a syntactic signal than a non-syntactic signal, $q/q_s = 1/2$, and if a fraction $p = 1/3$ of all noun verb combinations describe meaningful events, then at least an 18×18 system is required for syntactic communication to have any chance of evolving. Figure 4 shows the excellent agreement between our approximative analytical results and exact numerical computations.

The parameter p quantifies to what extent the perceived world has a compositional structure. A small p means that events often consist of unique pairings of objects and actions. The smaller the value of p , the harder it is for syntactic communication to evolve (the critical n in equation (3) is large).

Our results suggest that the crucial step that guided the transition from non-syntactic to syntactic communication was an increase in the number of relevant events that could be referred to. 'Relevant event' means there is a fitness contribution for communication about this event. As the number of such 'relevant communication topics' increased, natural selection could begin to favour syntactic

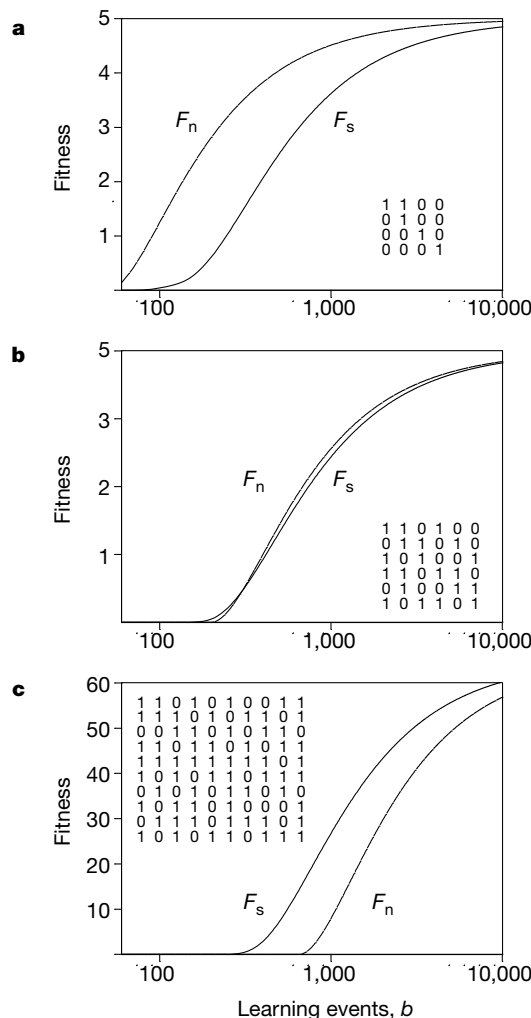


Figure 3 The fitness of non-syntactic and syntactic communication, F_n and F_s , as function of the total number of word learning events per individual, b , for three different choices of the event rate matrix, Γ . The entries of Γ are the numbers γ_{ij} which characterize the rates at which the various events occur. **a**, There are four objects and four actions. Each object is associated with a specific action; in addition object 1 also occurs with action 2. All possible events occur at the same rate. Thus the event rate matrix is a binary 4×4 matrix with 5 non-zero entries, $p = 5/16$. For b ranging from 50 to 10,000, F_n always exceeds F_s . **b**, There are six objects and six actions, the event rate matrix has 20 non-zero entries, $p = 5/9$. For values of b less than 400, syntactic communication has a higher fitness than non-syntactic communication. For values of b above 400, non-syntactic communication wins. Hence, for medium-sized systems the emergence of syntactic communication can be prevented by increasing the number of learning events per individual. **c**, There are 10 objects and 10 actions, 65 of 100 combinations occur, $p = 13/20$. In this case, syntactic communication wins for any choice of b . Each panel shows F_n and F_s as function of b and illustrates the chosen event rate matrix, Γ .

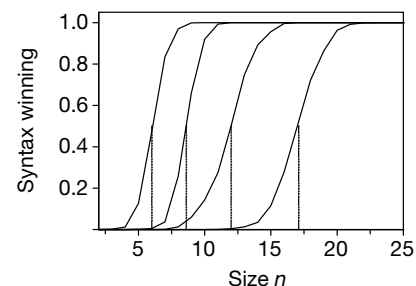


Figure 4 Numerical validation of the approximate threshold condition given by equation (3). If this inequality holds then syntax will be favoured by natural selection. There is communication about n objects and m actions. A fraction, p , of the nm object-action combinations refers to events that occur in the world. The parameters q and q_s denote, respectively, the probability of memorizing a non-syntactic and a syntactic signal after a single occurrence. We choose $n = m$, $b = 5,000$ and $q = 0.1$. The values of n range from 2 to 25. For each n we compute F_s and F_n for 1,000 randomly chosen event-rate matrices Γ , whose entries are 1 with probability p or 0 otherwise. The y-axis shows the fraction of times when F_s exceeds F_n . From left to right the curves correspond to the parameter values $p = 0.5$, $q_s = 0.1$; $p = 0.5$, $q_s = 0.07$; $p = 0.25$, $q_s = 0.1$; $p = 0.25$, $q_s = 0.07$. The straight lines indicate the predicted threshold values $n = 6, 8.6, 12, 17.1$.

communication and thereby lead to a language design where messages could be formulated that were not learned beforehand. Syntactic messages can encode new ideas or refer to extremely rare but important events. Our theory, however, does not suggest that syntactic communication is always at an advantage. Many animal species probably have a syntactic understanding of the world, but natural selection did not produce a syntactic communication system for these species because the number of relevant signals was below the threshold illustrated by equation (3). Presumably the increase in the number of relevant communication topics was caused by changes in the social structure²⁴ and interaction of those human ancestors who evolved syntactic communication. □

Methods

Suppose there are n words. Individuals are characterized by the subset of words they know. There are 2^n possibilities for the internal lexicon of an individual. Internal lexica are defined by bit strings: 1 means that the corresponding word is known; 0 means it is not. Let us enumerate them by $I = 0, \dots, \nu$ where $\nu = 2^n - 1$. The number I is the integer representation of the corresponding bit string. (For example, $I = 6$ represents the string 000...0110.) Denote by x_I the abundance of individuals with internal lexicon I . The population dynamics can be formulated as

$$\dot{x}_I = \delta_I - x_I + b \sum_{J=0}^{\nu} \sum_{K=0}^{\nu} (x_J x_K Q_{JKI} - x_I x_J Q_{IJK}) \quad (4)$$

where $I = 0, \dots, \nu$. We have $\delta_0 = 1$ and $\delta_I = 0$ otherwise; thus all individuals are born not knowing any of the words. Individuals die at a constant rate, which we set to 1, thereby defining a time scale. The quantities Q_{IJK} denote the probabilities that individual I learning from J will become K . Equation (4) is a general framework for the population dynamics of the lexical aspects of language. Assuming that in any one interaction between two individuals only a single new word can be acquired and that words are memorized independently of each other, we obtain the specific system described by equation (1).

Let us now assume that the world is made up of events E_{ij} consisting of objects i and actions j . The 'event rate matrix', Γ has the entries γ_{ij} which specify the relative rate of occurrence of event E_{ij} . Denote by ϕ_{ij} the frequency of occurrence of event E_{ij} . We have $\phi_{ij} = \gamma_{ij} / \sum_{ij} \gamma_{ij}$.

Non-syntactic communication uses words, W_{ij} , for events E_{ij} . The basic reproductive ratio of W_{ij} is given by $R(W_{ij}) = b\phi_{ij}$. If $R(W_{ij}) > 1$, the word W_{ij} will persist in the population, and at equilibrium the relative abundance of individuals who know this word is given by $x^*(W_{ij}) = 1 - 1/R(W_{ij})$.

The fitness contribution of a language can be formulated as the probability that two individuals know the correct word for a given event summed over all events and weighted with the rate of occurrence of these events. Hence, at equilibrium, the fitness of individuals using non-syntactic communication is given by

$$F_n = \sum_{ij} [x^*(W_{ij})^2] \gamma_{ij}. \quad (5)$$

For syntactic communication, we assume the event E_{ij} is described by the sentence $N_i V_j$. The population dynamics of individuals knowing both N_i and V_j are described by equation (2). The basic reproductive ratios are given by $R(N_i) = (b/2)q_i \phi(N_i)$ and $R(V_j) = (b/2)q_j \phi(V_j)$. The frequencies of occurrence are $\phi(N_i) = \sum_j \phi_{ij}$ and $\phi(V_j) = \sum_i \phi_{ij}$. If the basic reproductive ratios, $R(N_i)$ and $R(V_j)$, are greater than one, the equilibrium frequency of individuals who know both N_i and V_j is given by

$$x^*(N_i V_j) = \frac{[1 - 1/R(N_i)][1 - 1/R(V_j)]}{1 - 1/[R(N_i) + R(V_j)]} \quad (6)$$

At equilibrium, the fitness of syntactic communication is given by

$$F_s = \sum_{ij} [x^*(N_i V_j)]^2 \gamma_{ij}. \quad (7)$$

Assuming there are n objects and m actions that give rise to pnm meaningful events that all occur at the same frequency, we obtain $R(W_{ij}) = bq/(pnm)$. If we also assume that the combinations of objects and actions are arranged in a way that all nouns and all verbs, respectively, have about the same frequency, we can write $R(N_i) = bq_i/(2n)$ and $R(V_j) = bq_j/(2m)$. For the fitness values, we obtain $F_n = pnm[1 - 1/R(W_{ij})]^2$ and $F_s = pnm[1 - 1/R(N_i)]^2[1 - 1/R(V_j)]^2/[1 - 1/(R(N_i) + R(V_j))]^2$. Assuming that all involved basic reproductive ratios are well above one, we obtain for $F_s < F_n$ the condition $(m^2 n + mn^2)/(m^2 + mn + n^2) > (2q)/(pq_s)$. Defining the ratio, $\alpha = m/n$, we can rewrite this condition as

$$n > \frac{2q}{pq_s} \left(1 + \frac{1}{\alpha} - \frac{1}{\alpha + 1} \right). \quad (8)$$

Hence the size of the system has to exceed a critical threshold for syntax to be favoured by natural selection.

Received 1 November 1999; accepted 26 January 2000.

1. Von Frisch, K. *The Dance Language and Orientation of Bees* (Harvard Univ. Press, Cambridge, Massachusetts, 1967).
2. Marler, P. Birdsong and speech development: could there be parallels? *Am. Sci.* **58**, 669–673 (1970).
3. Wilson, E. O. Animal communication. *Sci. Am.* **227**, 52–60 (1972).
4. Gould, J. L. & Marler, P. Learning by instinct. *Sci. Am.* **256**, 74–85 (1987).
5. Burling, R. Primate calls, human language, and nonverbal communication. *Curr. Anthropol.* **34**, 25–53 (1989).
6. Cheney, D. L. & Seyfarth, R. M. *How Monkeys See the World: Inside the Mind of Another Species* (Chicago Univ. Press, 1990).
7. Hauser, M. *The Evolution of Communication* (MIT Press, Cambridge, Massachusetts, 1996).
8. Bickerton, D. *Species and Language* (Chicago Univ. Press, Chicago, 1990).
9. von Humboldt, W. *Linguistic Variability and Intellectual Development* (Pennsylvania Univ. Press, Philadelphia, 1972).
10. Chomsky, N. *Aspects of the Theory of Syntax* (MIT Press, Cambridge, Massachusetts, 1965).
11. Jackendoff, R. *The Architecture of the Language Faculty* (MIT Press, Cambridge, Massachusetts, 1997).
12. Pinker, S. & Bloom, P. Natural language and natural selection. *Behav. Brain Sci.* **13**, 707–784 (1990).
13. Pinker, S. *The Language Instinct*. (Harper Collins, New York, 1994).
14. Maynard Smith, J. & Szathmari, E. *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
15. Hurford, J. R., Studdert-Kennedy, M. & Knight, C. *Approaches to the Evolution of Language* (Cambridge Univ. Press, Cambridge, UK, 1998).
16. Nowak, M. A. & Krakauer, D. C. The evolution of language. *Proc. Natl Acad. Sci. USA* **96**, 8028–8033 (1999).
17. Chomsky, N. *Language and Mind* (Harcourt Brace Jovanovich, New York, 1972).
18. Chomsky, N. *Language and Problems of Knowledge: The Managua Lectures* (MIT Press, Cambridge, Massachusetts, 1988).
19. Premack, D. Gavage or the future history of the animal language controversy. *Cognition* **19**, 207–296 (1985).
20. Lieberman, P. *The Biology and Evolution of Language* (Harvard Univ. Press, Cambridge, Massachusetts, 1984).
21. Bates, E., Thal, D. & Marchman, V. in *Biological and Behavioural Determinants of Language Development* (eds Krasnegor et al.) (Erlbaum, Mahwah, NJ, 1991).
22. Zipf, G. K. *The Psychobiology of Language* (Houghton-Mifflin, Boston, 1935).
23. Miller, G. A. & Chomsky, N. *Handbook of Mathematical Psychology*. Vol. 2 (eds Luce, R. D., Bush, R. & Galanter, E.) 419–491 (Wiley, New York, 1963).
24. Dunbar, R. *Grooming, Gossip and the Evolution of Language* (Harvard Univ. Press, Cambridge, Massachusetts, 1996).

Acknowledgements

This work was supported by the Leon Levy and Shelby White Initiatives Fund, the Florence Gould Foundation, the J. Seward Johnson Sr Charitable Trusts, the Ambrose Monell Foundation, the National Science Foundation, the Wellcome Trust and the Alfred P. Sloan Foundation.

Correspondence and requests for materials should be addressed to M.A.N. (e-mail: nowak@ias.edu).

Glutamate spillover suppresses inhibition by activating presynaptic mGluRs

Simon J. Mitchell & R. Angus Silver

Department of Physiology, University College London, Gower Street, London WC1E 6BT, UK

Metabotropic glutamate receptors (mGluRs) found on synaptic terminals throughout the brain are thought to be important in modulating neurotransmission^{1,2}. Activation of mGluRs by synaptically released glutamate depresses glutamate release from excitatory terminals^{3–5} but the physiological role of mGluRs on inhibitory terminals is unclear. We have investigated activation of mGluRs on inhibitory terminals within the cerebellar glomerulus, a structure in which GABA (γ -aminobutyric acid)-releasing inhibitory terminals and glutamatergic excitatory terminals are in close apposition and make axo-dendritic synapses onto granule cells⁶. Here we show that 'spillover' of glutamate, which is released from excitatory mossy fibres, inhibits GABA release from Golgi cell terminals by activating presynaptic mGluRs under physiological conditions. The magnitude of the depression of the

inhibitory postsynaptic current is dependent on the frequency of mossy fibre stimulation, reaching 50% at 100 Hz. Furthermore, the duration of inhibitory postsynaptic current depression mirrors the time course of mossy fibre activity. Our results establish that mGluRs on inhibitory interneuron axons⁷ sense the activity of neighbouring excitatory synapses. This heterosynaptic mechanism is likely to boost the efficacy of active excitatory fibres by locally reducing the level of inhibition.

Immunohistochemical^{17–10} and electrophysiological studies^{11,12} have shown that mGluRs 1/5 and 2/3 (groups I and II) are present on cerebellar Golgi cells. We therefore characterized the effect of activating mGluRs on evoked inhibitory postsynaptic currents (IPSCs) in granule cells using a broad-spectrum agonist. Application of ACPD ((±)-1-A-aminocyclopentane-*trans*-1,3-dicarboxylic acid; 100 μ M) reduced the amplitude of the evoked IPSC substantially at a low stimulation frequency ($75 \pm 5\%$ at 0.2 Hz, $P < 0.001$,

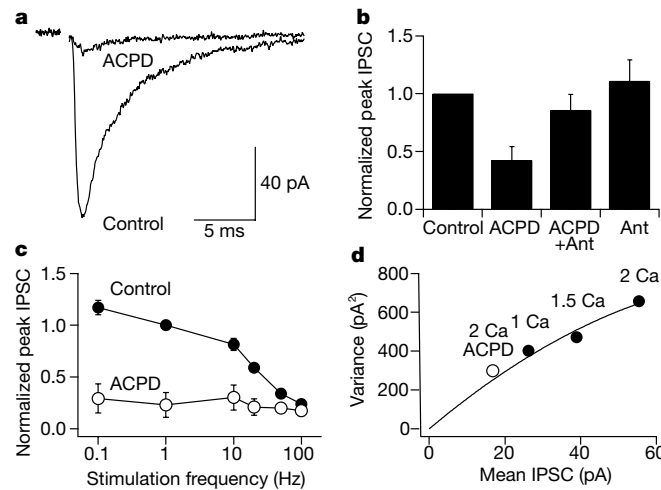


Figure 1 Activation of presynaptic mGluRs depresses IPSCs in granule cells. **a**, Average IPSCs evoked by Golgi axon stimulation at 0.2 Hz in control solution and 100 μ M ACPD. **b**, Pooled results ($n = 4$) showing effect of 20 μ M ACPD and its block by the antagonists (Ant) MCPG (1 mM) and CPPG (300 μ M). **c**, Steady-state frequency-response of evoked IPSCs normalized to 1 Hz value (filled circles); relationship in 100 μ M ACPD (open circles);

$n = 7$; 20 Hz and 50 Hz, $n = 6$). **d**, Variance–mean relationship for IPSCs recorded at 5 Hz at different release probabilities imposed with different external Ca^{2+} / Mg^{2+} solutions (filled circles). A parabolic fit gave a weighted mean quantal size of 357 pS for this cell. This remained unchanged during the 100 μ M ACPD application.

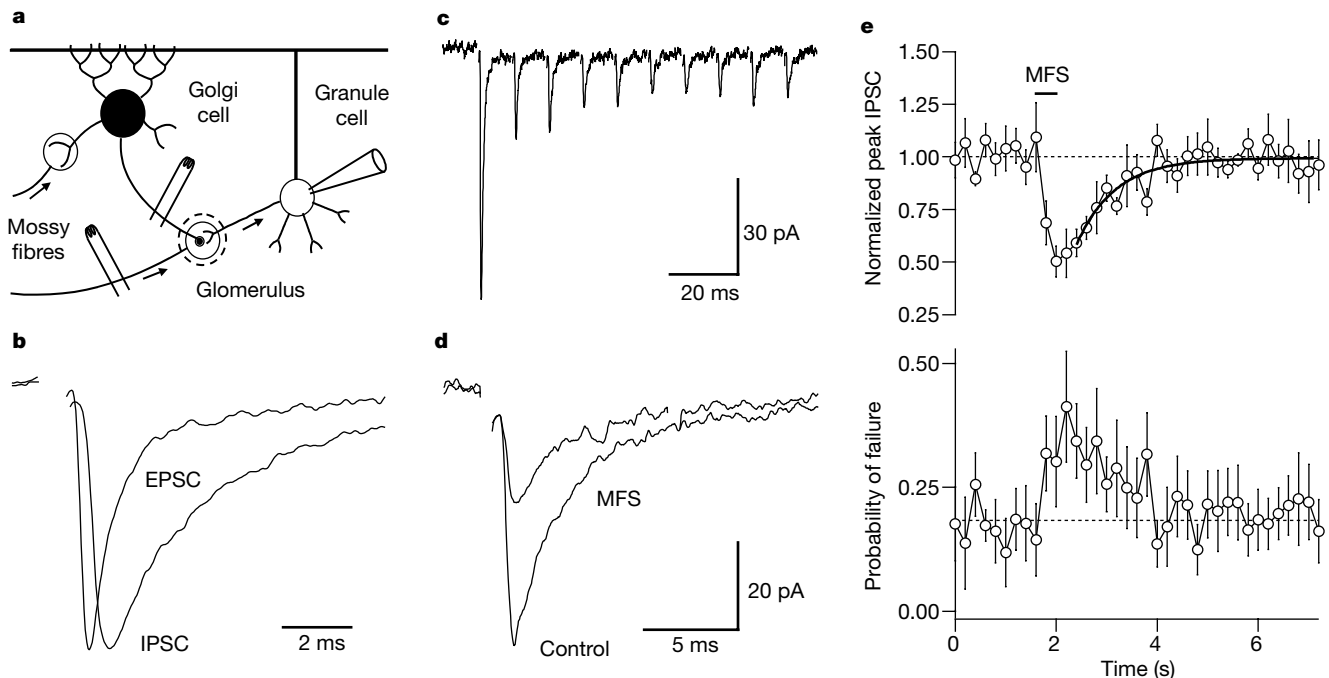


Figure 2 Mossy fibre stimulation depresses GABA release from Golgi cells in the absence of fast excitatory transmission. **a**, Recording configuration. Whole-cell patch-clamped granule cell and independent stimulation of inhibitory and excitatory inputs. **b**, Normalized EPSC and IPSC exhibit different kinetics. **c**, Averaged non-NMDA EPSCs depress rapidly at 100 Hz. **d**, Average IPSC waveforms from control period and during MFS. **e**, Top panel,

normalized IPSC amplitude during 5-Hz Golgi cell stimulation in ionotropic glutamate receptor antagonists. Bar indicates timing of 0.5-s burst of 100-Hz MFS. Bold line shows single exponential function fit to IPSC recovery ($\tau = 770$ ms). Bottom panel, probability of failure increases during IPSC depression.

$n = 6$; Fig. 1a) without altering the input resistance. As the group III agonist L-AP4 (L(+)-2-amino-4-phosphonobutyric acid) also depressed evoked IPSCs ($n = 3$), we used a broad-spectrum combination of mGluR antagonists, MCPG ((S)-methyl-4-carboxyphenylglycine) and CPPG ((RS)- α -cyclopropyl-4-phosphonophenylglycine), to block activation (Fig. 1b).

We examined the effects of mGluR activation over a range of stimulation frequencies because Golgi cells fire at up to 100 Hz *in vivo*¹³. Under control conditions, the steady-state amplitude of the evoked IPSC declined as Golgi cell firing was raised above the spontaneous firing rate (8 ± 2 Hz; $n = 8$) by stimulation (Fig. 1c). In the presence of ACPD, however, the frequency dependence of the evoked IPSC was eliminated because of a more marked depression at lower frequencies. Application of multiple-probability fluctuation analysis^{14,15} indicated that the quantal size was unchanged during IPSC depression ($99 \pm 9\%$, $P = 0.9$, $n = 3$; Fig. 1d), confirming that ACPD reduces the probability of GABA release. These effects are probably mediated predominantly by mGluR2 and/or mGluR3, which are present on Golgi cell presynaptic terminals⁷, rather than mGluR5 which is located on the somatodendritic compartment¹⁰. In addition to direct inhibition of GABA release, ACPD also increased the spontaneous Golgi cell firing rate (to 13 ± 2 Hz, $P = 0.003$, $n = 8$), as previously reported for both cerebellar molecular layer and hippocampal interneurons^{16,17}; however, the frequency-dependent depression associated with this increase in firing represented only a small fraction ($\sim 30\%$) of the ACPD-mediated IPSC depression.

Despite considerable interest in presynaptic receptors, there is no direct evidence that mGluRs found on inhibitory terminals are activated by spillover of glutamate under physiological conditions. Several studies suggest that GABA and glutamate can spillover in the cerebellar glomerulus and activate postsynaptic receptors at neighbouring synapses or in the extrasynaptic membrane^{18–22}. We therefore determined whether glutamate released from mossy fibres can activate mGluRs on Golgi cells. We examined this at physiological

temperature ($\sim 37^\circ\text{C}$) when rates of glutamate uptake and diffusion are maximal, because these processes will determine the distance over which synaptically released glutamate can activate mGluRs. Our experimental protocol involved stimulating independent mossy fibre and Golgi cell inputs onto a granule cell (Fig. 2a). Evoked excitatory postsynaptic currents (EPSCs) and monosynaptic IPSCs were distinguished by their kinetics and reversal potential (Figs 2b and 3a). The slice was then perfused with ionotropic glutamate receptor antagonists (see Methods) to prevent synaptic activation of Golgi cells and granule cells during mossy fibre stimulation (MFS).

The efficacy of the inhibitory Golgi cell input was monitored from the amplitude of evoked IPSCs during a 5-Hz stimulation train. Several seconds into the IPSC train, the mossy fibre input was stimulated at 100 Hz, a frequency within the range measured *in vivo*¹³ (this protocol caused depression of the non-NMDA (N-methyl-D-aspartate) EPSC prior to application of antagonists; see Fig. 2c). A marked depression in the IPSC amplitude (Fig. 2d, e) was observed during MFS in 48% (16 out of 33 cells) of recordings. If a single mossy fibre is activated, a 100% success rate would require that all four glomeruli from which an individual granule cell receives input²³ are connected to the same Golgi cell axon and that the axon remains intact during slice preparation. The large fraction of responsive cells that we observed suggests that most glomeruli, if not all, exhibit IPSC depression that is induced by mossy fibre activity. The average time course of IPSC depression during MFS at 100 Hz for 500 ms is shown in Fig. 2e. At the peak of the response, the IPSC was depressed by $50 \pm 7\%$ ($P = 0.001$, $n = 6$; Fig. 2d, e) and the IPSC recovery waveform was fit with a single exponential function with a time constant of 770 ms. During IPSC depression, the probability of IPSC failures doubled (from 0.18 ± 0.06 to 0.34 ± 0.07 , $P = 0.03$, $n = 6$; Fig. 2e), indicating a reduction in the probability of GABA release.

If the Golgi cell axon was directly activated by the MFS electrode or the spontaneous Golgi cell firing rate was increased during MFS, the IPSC depression might have been caused by 'depletion' of inhibitory terminals rather than by activation of presynaptic receptors; however, examination of the traces during MFS showed that no synaptic currents were evoked at 100 Hz after blocking ionotropic glutamate receptors (compare Fig. 3a with b, c) and that the frequency of spontaneous IPSCs remained the same ($P = 0.92$, $n = 4$). In contrast to evoked IPSCs, the amplitude of spontaneous IPSCs remained unchanged during MFS ($P = 0.31$, $n = 4$), suggesting that IPSC depression occurred only at glomeruli activated by MFS. We tested whether mGluRs were involved in evoked IPSC depression by applying 300 μM CPPG and 1 mM MCPG. In the presence of these mGluR antagonists the MFS-induced depression

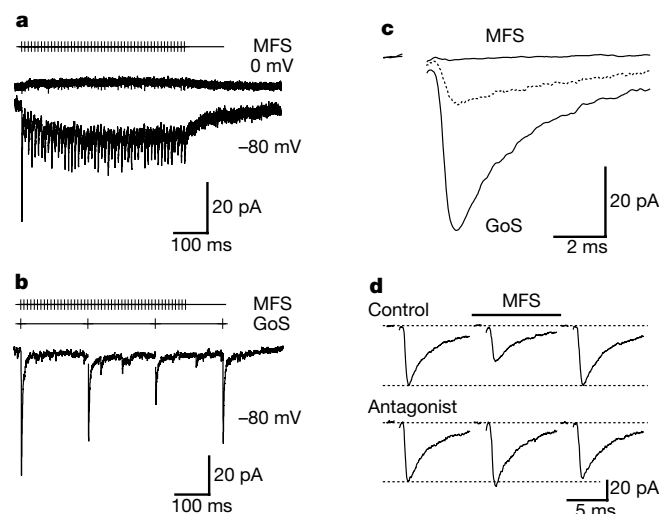


Figure 3 Metabotropic glutamate receptors mediate MFS-induced IPSC depression. **a**, 100-Hz MFS (timing indicated by ticks) at 0 mV and -80 mV without glutamate receptor blockers. The lack of evoked IPSCs, which had a reversal potential of -24 mV, shows that MFS was independent of inhibitory input. **b**, Simultaneous 5-Hz Golgi cell stimulation (GoS) and 100-Hz MFS in ionotropic glutamate receptor blockers. The IPSC depresses during the 0.5-s MFS, whereas no currents were associated with the MFS (small currents throughout are averaged spontaneous IPSCs). **c**, Stimulus-averaged currents. GoS indicates IPSC at 5 Hz and MFS shows that IPSCs were not evoked during 100-Hz MFS. Dashed line shows, for comparison, 100-Hz GoS for 0.5 s. **d**, Averaged IPSCs before, during and after 100-Hz MFS for control and in the presence of mGluR antagonists MCPG (1 mM) and CPPG (300 μM).

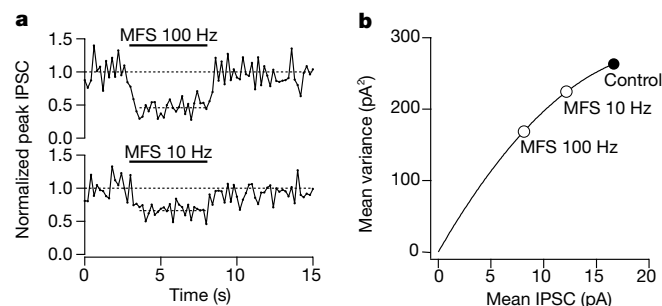


Figure 4 IPSC depression depends on duration and frequency of MFS. **a**, Normalized IPSC amplitude during Golgi cell stimulation at 5 Hz in ionotropic glutamate receptor antagonists. Bars indicate timing of prolonged MFS at 100 Hz and 10 Hz. Dashed lines indicate mean values for the baseline and during MFS. **b**, Variance-mean relationship for IPSCs evoked at 5 Hz for control (filled circle) and during prolonged MFS at 10 Hz and 100 Hz (open circles). The parabolic fit gave a weighted mean quantal size of 456 pS for this cell.

of the IPSC was blocked ($106 \pm 8\%$ of control, $P = 0.77$, $n = 4$; Fig. 3d), providing direct evidence that IPSC depression was caused by mGluR activation. Consistent with mGluR activation by glutamate, application of the uptake blocker PDC (L-trans-pyrrolidine-2,4-dicarboxylic acid), which raises extracellular glutamate through heteroexchange, produced a substantial tonic IPSC depression ($51 \pm 12\%$ at $1 \mu\text{M}$). Mossy fibre stimulation during PDC application consistently depressed the IPSC further (to $64 \pm 11\%$ at 10 Hz , $n = 3$).

To understand the physiological role of presynaptic mGluRs on Golgi cells, we examined the effects of prolonged activity and lower MFS frequencies on the IPSC. The IPSC amplitude remained depressed during MFS for 5 s at 100 Hz (Fig. 4a), showing no significant accommodation over this timescale. The IPSC depression reached a steady-state level of $54 \pm 11\%$ ($n = 4$) within $\sim 0.5 \text{ s}$ of the onset of MFS. Sustained depression of the IPSC also occurred when the same mossy fibres were stimulated at lower frequencies (Fig. 4a); however, at 10 Hz MFS the extent of the depression was significantly smaller ($34 \pm 10\%$, $P = 0.03$, $n = 4$). Synaptic variance–mean relationships constructed from IPSCs during sustained MFS at 10 Hz and 100 Hz (Fig. 4b) gave an estimate of quantal size ($350 \pm 60 \text{ pS}$, $n = 4$; Fig. 1d) that was similar to that obtained with multiple-probability fluctuation analysis when release probability was reduced with different $\text{Ca}^{2+}/\text{Mg}^{2+}$ solutions ($410 \pm 60 \text{ pS}$, $n = 10$, $P = 0.3$, unpaired t -test), indicating that sustained IPSC depression was also mediated presynaptically. These results show that the magnitude and duration of IPSC depression reflect the frequency and duration of mossy fibre activity.

Our results show that during excitatory mossy-fibre activity glutamate spills over onto inhibitory Golgi cell terminals and depresses GABA release by activating mGluRs on the sub-second timescale. Metabotropic glutamate receptor activation is likely to be locally restricted to the axon terminals within active glomeruli because the concentration of glutamate falls rapidly with distance from the release site²⁴ and glomeruli are partially encapsulated by glia at this age²⁵. Such local activation of presynaptic mGluRs by glutamate spillover allows modulation of inhibition at individual synaptic inputs. A related mechanism involving reciprocal synapses has been proposed in the olfactory bulb²⁶. Communication between excitatory and inhibitory synapses provides a local feedforward disinhibitory mechanism that contrasts with depolarization-induced suppression of inhibition, an mGluR-mediated feedback mechanism that depends on postsynaptic activity through a retrograde messenger^{27,28}. The spillover-mediated disinhibition of granule cells will boost the efficacy of active mossy fibres, thereby enhancing contrast between inputs firing at different rates. As suppression of inhibition showed little accommodation, presynaptic mGluRs on Golgi cells are likely to be important for processing information encoded as tonic firing rate, such as joint angle¹³. The firing rate of Golgi cells *in vivo*¹³ reflects the level of excitation in the regional population of mossy fibres and granule cells^{23,29,30}. As inhibition of GABA release by mGluRs is reduced at higher Golgi cell firing rates (Fig. 1c), contrast enhancement by glutamate spillover is likely to be most effective when the mean firing rate of the mossy fibre population is low. □

Methods

Sagittal slices of cerebellum ($200\text{--}250 \mu\text{m}$) were prepared from 12–13-day-old Sprague-Dawley rats¹⁸. Recording solution contained (in mM): 125 NaCl, 2.5 KCl, 2 CaCl_2 , 1 MgCl_2 , 1.25 NaH_2PO_4 , 26 NaHCO_3 and 25 glucose (pH 7.3 when bubbled with 95% O_2 and 5% CO_2 , 310 mOsm). All recordings were made at $34\text{--}39^\circ\text{C}$. $10 \mu\text{M}$ D-2-amino-5-phosphonopentanoic acid (D-AP5), $20 \mu\text{M}$ 7-chlorokynurenic acid, $2 \mu\text{M}$ 2,3-dioxo-6-nitro-1,2,3,4-tetrahydrobenzo [f]quinoxaline-7-sulphonamide (NBQX) and $0.5 \mu\text{M}$ strychnine were added to the recording solution to block NMDA, non-NMDA and glycine receptors. $20\text{--}100 \mu\text{M}$ ACPD, $50 \mu\text{M}$ L-AP4, 1 mM MCPG and $300 \mu\text{M}$ CPPG were added to agonize or antagonize mGluR receptors. $1 \mu\text{M}$ PDC was added to block glutamate uptake.

Recordings were made using an Axopatch 200B amplifier (Axon Instruments), $5\text{--}10 \text{ M}\Omega$ micropipettes made from borosilicate glass and an internal solution containing (in mM) 140 CsCl, 10 Cs-HEPES, 10 Cs-EGTA, 2 NaCl and 2 Mg-ATP (pH 7.3;

290 mOsm) to block postsynaptic K^+ conductances. In experiments where excitatory and inhibitory inputs were activated, a pipette solution containing (in mM) 75 Cs-methylsulphonate, 50 CsCl, 10 Cs-HEPES, 10 Cs-EGTA, 2 NaCl and 2 Mg-ATP (chloride reversal potential of -24 mV) was used allowing EPSCs and IPSCs to be viewed separately. Cell-attached recordings from Golgi cells and stimulation pipettes were made using patch electrodes containing extracellular solution. Golgi cell and mossy fibre axons were stimulated with isolated constant voltage devices (Digitimer DS2). A single fibre was identified on the basis of an all-or-none postsynaptic current elicited by a stimulus of graded intensity. The stimulation voltage was set above threshold to ensure reliable fibre stimulation¹⁸. Data was acquired at $10\text{--}25 \text{ kHz}$ using Axograph 4 software and an Instrutech ITC-18 interface after low-pass filtering at $2\text{--}5 \text{ kHz}$ (Bessel filter). Time stability of synaptic currents was assessed using Spearman rank order correlation analysis. IPSC detection threshold for failure analysis was two standard deviations of the background noise. Analysis was confirmed by visual inspection of averaged failures. Means are expressed $\pm \text{s.e.m.}$; groups were compared using a paired Student's t -test except where stated.

Received 15 November 1999; accepted 23 February 2000.

- Nakanishi, S. Metabotropic glutamate receptors: synaptic transmission, modulation and plasticity. *Neuron* **13**, 1031–1037 (1994).
- Conn, P. J. & Pin, J.-P. Pharmacology and functions of metabotropic glutamate receptors. *Annu. Rev. Pharmacol. Toxicol.* **37**, 205–237 (1997).
- Scanziani, M., Salin, P. A., Vogt, K. E., Malenka, R. C. & Nicoll, R. A. Use-dependent increases in glutamate concentration activate presynaptic metabotropic glutamate receptors. *Nature* **385**, 630–634 (1997).
- von Gersdorff, H., Schneggenburger, R., Wiess, S. & Neher, E. Presynaptic depression at a calyx synapse: the small contribution of metabotropic glutamate receptors. *J. Neurosci.* **17**, 8137–8146 (1997).
- Vogt, K. E. & Nicoll, R. A. Glutamate and γ -aminobutyric acid mediate a heterosynaptic depression at mossy fiber synapses in the hippocampus. *Proc. Natl Acad. Sci. USA* **96**, 1118–1122 (1999).
- Jakab, R. L. & Hámori, J. Quantitative morphology and synaptology of cerebellar glomeruli in the rat. *Anat. Embryol.* **179**, 81–88 (1988).
- Ohishi, H. *et al.* Immunohistochemical localization of metabotropic glutamate receptors, mGluR2 and mGluR3, in rat cerebellar cortex. *Neuron* **13**, 55–66 (1994).
- Baude, A. *et al.* The metabotropic glutamate receptor (mGluR1 α) is concentrated at perisynaptic membrane of neuronal subpopulations as detected by immunogold reaction. *Neuron* **11**, 771–787 (1993).
- Neki, A. *et al.* Metabotropic glutamate receptors mGluR2 and mGluR5 are expressed in two non-overlapping populations of Golgi cells in the rat cerebellum. *Neuroscience* **75**, 815–826 (1996).
- Jaarsma, D., Dino, M. R., Ohishi, H., Shigemoto, R. & Mugnaini, E. Metabotropic glutamate receptors are associated with non-synaptic appendages of unipolar brush cells in rat cerebellar cortex and cochlear nuclear complex. *J. Neurocytol.* **27**, 303–327 (1998).
- Knoflach, F. & Kemp, J. A. Metabotropic glutamate group II receptors activate a G protein-coupled inwardly rectifying K^+ current in neurones of the rat cerebellum. *J. Physiol. (Lond.)* **509**, 347–354 (1998).
- Vetter, P., Garthwaite, J. & Batchelor, A. M. Regulation of synaptic transmission in the mossy fibre-granule cell pathway of rat cerebellum by metabotropic glutamate receptors. *Neuropharmacology* **38**, 805–815 (1999).
- van Kan, P. L., Gibson, A. R. & Houk, J. C. Movement-related inputs to intermediate cerebellum of monkeys. *J. Neurophysiol.* **69**, 74–94 (1993).
- Silver, R. A., Momiyama, A. & Cull-Candy, S. G. Locus of frequency-dependent depression identified with multiple-probability fluctuation analysis at rat climbing fibre-Purkinje cell synapses. *J. Physiol. (Lond.)* **510**, 881–902 (1998).
- Clements, J. D. & Silver, R. A. Unveiling synaptic plasticity: a new graphical and analytical approach. *Trends Neurosci.* **23**, 105–113 (2000).
- Llano, I. & Marty, A. Presynaptic metabotropic glutamatergic regulation of inhibitory synapses in rat cerebellar slices. *J. Physiol. (Lond.)* **486**, 163–176 (1995).
- Poncer, J. C., Shinokaki, H. & Miles, R. Dual modulation of synaptic inhibition by distinct metabotropic glutamate receptors in the rat hippocampus. *J. Physiol. (Lond.)* **485**, 121–134 (1995).
- Silver, R. A., Cull-Candy, S. G. & Takahashi, T. Non-NMDA glutamate receptor occupancy and open probability at a rat cerebellar synapse with single and multiple release sites. *J. Physiol. (Lond.)* **494**, 231–250 (1996).
- Tia, S., Wang, J. F., Kotchabhakdi, N. & Vicini, S. Developmental changes of inhibitory synaptic currents in cerebellar granule neurons: role of GABA $_A$ receptor $\alpha 6$ subunit. *J. Neurosci.* **16**, 3630–3640 (1996).
- Brickley, S. G., Cull-Candy, S. G. & Farrant, M. Development of a tonic form of synaptic inhibition in rat cerebellar granule cells resulting from persistent activation of GABA $_A$ receptors. *J. Physiol. (Lond.)* **497**, 753–759 (1996).
- Wall, M. J. & Usowicz, M. M. Development of action potential-dependent and independent spontaneous GABA $_A$ receptor-mediated currents in granule cells of postnatal rat cerebellum. *Eur. J. Neurosci.* **9**, 533–548 (1997).
- Rossi, D. J. & Hamann, M. Spillover-mediated transmission at inhibitory synapses promoted by high affinity $\alpha 6$ subunit GABA $_A$ receptors and glomerular geometry. *Neuron* **20**, 783–795 (1998).
- Eccles, J. C., Ito, M. & Szentagothai, J. *The Cerebellum as a Neuronal Machine* (Springer, Germany, 1967).
- Barbour, B. & Häusser, M. Intersynaptic diffusion of neurotransmitter. *Trends Neurosci.* **20**, 377–384 (1997).
- Hámori, J. & Somogyi, J. Differentiation of cerebellar mossy fiber synapses in the rat: a quantitative electron microscope study. *J. Comp. Neurol.* **220**, 365–377 (1983).
- Hayashi, Y., Momiyama, A., Takahashi, T., Ohishi, H., Ogawa-Meguro, R., Shigemoto, R., Mizuno, N. & Nakanishi, S. Role of a metabotropic glutamate receptor in synaptic modulation in the accessory olfactory bulb. *Nature* **366**, 687–690 (1993).
- Glitsch, M., Llano, I. & Marty, A. Morishita Glutamate as a candidate retrograde messenger at interneurone-Purkinje cell synapses of rat cerebellum. *J. Physiol. (Lond.)* **497**, 531–537 (1996).

28. Morishita, W., Kirov, S. A. & Alger, B. E. Evidence for metabotropic glutamate receptor activation in the induction of depolarization-induced suppression of inhibition in hippocampal CA1. *J. Neurosci.* **18**, 4870–4882 (1998).
29. Marr, D. A theory of cerebellar cortex. *J. Physiol. (Lond.)* **202**, 437–470 (1969).
30. Ito, M. in *The Cerebellum and Neural Control*. (Raven, New York, 1984).

Acknowledgements

This work was supported by The Wellcome Trust, the European Union and the Medical Research Council (Research Studentship to S.J.M.). We thank A. Momiyama and T. Takahashi for helpful discussion, J. Clements for providing Axograph and D. Attwell, M. Farrant, M. Häusser, B. Katz and T. Takahashi for comments on the manuscript.

Correspondence and requests for materials should be addressed to R.A.S. (e-mail: a.silver@ucl.ac.uk).

Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491

J. Parkhill*, M. Achtman†, K. D. James*, S. D. Bentley*, C. Churcher*, S. R. Klee†, G. Morelli†, D. Basham*, D. Brown*, T. Chillingworth*, R. M. Davies*, P. Davis*, K. Devlin*, T. Feltwell*, N. Hamlin*, S. Holroyd*, K. Jagels*, S. Leather*, S. Moule*, K. Mungall*, M. A. Quail*, M.-A. Rajandream*, K. M. Rutherford*, M. Simmonds*, J. Skelton*, S. Whitehead*, B. G. Spratt‡ & B. G. Barrell*

* The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

† Max-Planck Institut für molekulare Genetik, Ihnestr. 73, D-14195 Berlin, Germany

‡ Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3FY, UK

Neisseria meningitidis causes bacterial meningitis and is therefore responsible for considerable morbidity and mortality in both the developed and the developing world. Meningococci are opportunistic pathogens that colonize the nasopharynx and oropharynx of asymptomatic carriers. For reasons that are still mostly unknown, they occasionally gain access to the blood, and subsequently to the cerebrospinal fluid, to cause septicaemia and meningitis. *N. meningitidis* strains are divided into a number of serogroups on the basis of the immunochemistry of their capsular polysaccharides; serogroup A strains are responsible for major epidemics and pandemics of meningococcal disease, and therefore most of the morbidity and mortality associated with this disease. Here we have determined the complete genome sequence of a serogroup A strain of *Neisseria meningitidis*, Z2491 (ref. 1). The sequence is 2,184,406 base pairs in length, with an overall G+C content of 51.8%, and contains 2,121 predicted coding sequences. The most notable feature of the genome is the presence of many hundreds of repetitive elements, ranging from short repeats, positioned either singly or in large multiple arrays, to insertion sequences and gene duplications of one kilobase or more. Many of these repeats appear to be involved in genome fluidity and antigenic variation in this important human pathogen.

The genome encodes complete sets of enzymes for glycolysis (apart from *fruK* and *pfkA*), gluconeogenesis, the pentose-phosphate and Entner–Doudoroff pathways, the pyruvate dehydrogenase complex and the trichloroacetic acid cycle. In addition to the aerobic respiration genes, there are also both nitrite (*aniA*) and nitrate (NMA1886) reductases; some capability for fermentation is

also apparent. *N. meningitidis* appears to be capable of *de novo* synthesis of most of the amino acids (with the exception of asparagine and methionine) and purine and pyrimidine nucleotides; however, the pathways for production of folic acid, molybdopterin, pantothenate and pyridoxine are incomplete. All of the aminoacyl transfer RNA synthetases are present except tRNA^{Asn}. As *N. meningitidis* encodes both glutamyl tRNA synthetase and the three-subunit Glu-tRNA^{Gln} transamidase, the latter may be responsible for the production of tRNA^{Asn} by the transamidation of tRNA^{Asp}.

In addition to the 2,121 predicted coding sequences (CDSs), there are 4 copies of a 16S–23S–5S ribosomal RNA operon, 58 tRNAs, 1 tmRNA (10Sa RNA) and the RNA component of RNAase P. The average gene length is 877 base pairs (bp), at a density of 1 per 1.03 kilobases (kb). The overall coding density is 82.9%, higher than *Rickettsia*³, but lower than most other sequenced bacteria. Base 1 of the sequence was chosen to correspond roughly with 12 o'clock on the published map¹; the origin of replication (as indicated by the bias towards G on the leading strand⁴) is around 247,600 bp. Intriguingly, although the GC bias is similar in both replicons, there is a clear bias (60.2%) of CDSs towards the leading strand in only one replicon (anticlockwise in Fig. 1), with no coding bias at all on the opposite replicon. The genome contains at least 56 pseudogenes, of which 17 are remnants of insertion sequence (IS) elements, and there are at least 5 complete or partial prophages (pnm1–5).

The G+C content of the genome is extremely variable, with at least 60 coding regions (nearly 5% in total) having a significantly lower G+C content, ranging in size from 224 bp to 11.3 kb and averaging 1.8 kb (see Supplementary Information). These regions may represent recently acquired DNA, and given the natural competence of *Neisseria* species⁵, this is perhaps unsurprising. Although most genes in these regions have no database matches, or are conserved hypothetical proteins, nearly one-third encode

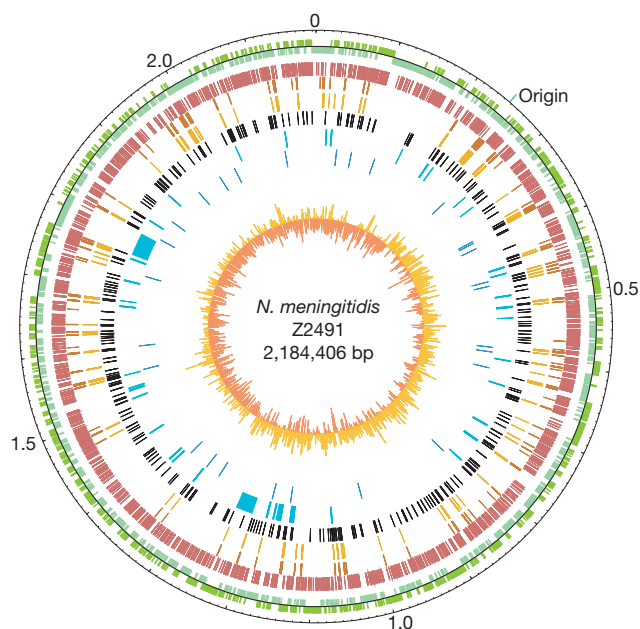


Figure 1 Circular representation of the *N. meningitidis* Z2491 genome. The concentric circles show, reading inwards: the scale in megabases, with the origin of replication indicated; predicted coding sequences clockwise (dark green) and anti-clockwise (light green); neisserial uptake sequences (red); dRS3 sequences (dark orange); RS elements (light orange); dispersed repeats (Correia, ATR, REP2-5; black); IS elements and phage (narrow ticks and wide bars respectively; turquoise) and tandem repeats (dark blue). The inner histogram shows plot of (G-C)/(G+C) with values greater than zero in yellow and less than zero in orange. Figure generated with LASERGENE software (DNASStar).

proteins that are likely to be on the surface of the cell, or are responsible for the production of surface structures. Prominent among these are the well characterized CDSs of the serogroup A capsule cassette (*sacA-D*) which produce a serogroup-A-specific sugar residue in the capsular polysaccharide⁶; these are located in an extremely well defined region of only 29.7% G+C. Also notable are a single glycosyl-transferase (NMA0832), similar to capsule biosynthesis genes in other bacteria, and a pilin homologue (NMA0424). Some surface-exposed protein-encoding genes are each immediately followed by large regions of A+T-rich DNA encoding numerous small hypothetical proteins; these include the three *mafB* family genes and NMA1083. Eight A+T-rich regions encode DNA restriction/modification enzymes. Some caution is required in interpreting a bias in G+C composition as necessarily indicating recently acquired DNA; the largest region of A+T-rich DNA (41.7% G+C over 11.3 kb) encompasses much of the ribosomal protein operon.

One of the most striking and unique characteristics of the Z2491 genome is the abundance and diversity of repetitive DNA (Fig. 1). The most obvious example of this abundance is the Neisserial uptake sequence, which is involved in the recognition and uptake of DNA from the environment⁷. There are nearly two thousand copies of the 10-bp uptake sequence, which either occurs alone or in inverted repeats as part of a transcriptional terminator⁷. Other repetitive sequence elements in the *N. meningitidis* genome are concentrated within intergenic repeat arrays of 200–2,700 bp. These repeat arrays are composed of several different repeat types (Fig. 2 and Table 1), the most abundant of which are the 'neisserial intergenic mosaic elements' (NIMEs), which comprise repeat units of ~50–150 bp (RS elements), each flanked by 20-bp inverted repeats (dRS3 elements); there are between 1 and 60 copies of 117 different families of RS element.

Also present in the repeat arrays are larger units, which are also found in isolation, including the 'Correia elements' (CEs, 156-bp sequences bounded by 26-bp inverted repeats) that have been previously described in both *N. meningitidis* and *Neisseria gonorrhoeae*⁸. Almost one-third of the 257 CEs share a common 51-bp internal deletion, which presumably has not affected the ability of the element to be mobilized, and 29 have lost one or both

of their terminal inverted repeats. Correia elements in repeat arrays are sometimes flanked by additional conserved sequences upstream and downstream (Fig. 2). A number of both complete and internally deleted CEs occur embedded within other sequence features, including genes, RS elements, other CEs and ISs. The pseudogenes NMA0059 (a modification methylase) and NMA0530 (a hypothetical protein) both appear to have been truncated or interrupted by CE insertion. Two repeat types are dispersed over the genome at low frequency, both alone and within repeat arrays: 19 copies of a 183-bp A+T-rich (30% G+C) repeat (ATR) whose ends form an imperfect 35-bp inverted repeat and 26 copies of a 120–150-bp repeat (REP 2). REP 2 occurs immediately upstream of 16 CDSs and contains a ribosome-binding-site-like conserved AAGGA motif within 5–13 bp of the predicted start codon. It seems likely that sequences within this repeat would exert some effect on the expression of these genes. The repeat arrays appear to be a target for the insertion of the mobile elements in the genome, with almost half of the 43 complete and partial copies of IS1016, IS1106 and IS1655 being integrated in or near to a repeat array.

Given the natural competence of neisserial species, the presence of such arrays could encourage sequence variation by acting as a target for a specific recombinase, thereby increasing the rate of horizontal gene transfer at the flanking locus, resulting in rapid allelic replacement or the creation of new mosaic alleles. A previously characterized example of enhanced recombinational exchange involves the rearrangements responsible for the modulation of *pilE* expression. *PilE* is the main structural pilin component and is expressed from the intact *pilE* gene. In *N. gonorrhoeae*, 'silent' pilin genes (*pilS*) exist, which contain only the 3' region of the gene without a start codon or transcriptional signals⁹. Coding sequences from these silent regions can be recombined into the expression locus, in a *recA/Q/O*-dependent manner, changing the expressed *PilE* sequence¹⁰. Models put forward for this process suggest that the gene conversion may be intergenomic¹¹. In the *N. meningitidis* Z2491 sequence, the *pilE* gene is surrounded by repeat arrays consisting of NIMEs and a CE (Fig. 3a). Immediately upstream of the *pilE* gene are eight *pilS* loci, each without an 5' end and intimately bounded by arrays of NIMEs, suggesting that NIME sequences may be involved in *pilS/E* recombination. The *N. gonorrhoeae* Sma/Cla repeat, implicated in enhancing *pilE/S* recombination in *N. gonorrhoeae*¹², exists as a single copy in *N. meningitidis* Z2491, immediately downstream of the *pilE* gene.

Our analysis of the occurrence of repeat arrays suggests that antigenic variation mediated by gene conversion, similar to that of *pilE*, may occur widely in the *N. meningitidis* genome. In *N. meningitidis* Z2491, most shorter repeat arrays are associated with genes directly or indirectly related to cell-surface functions (including pilins, substrate-binding proteins and transporter components), and the larger repeat arrays are exclusively associated with such genes (Fig. 4; and Table 2, Supplementary Information). The

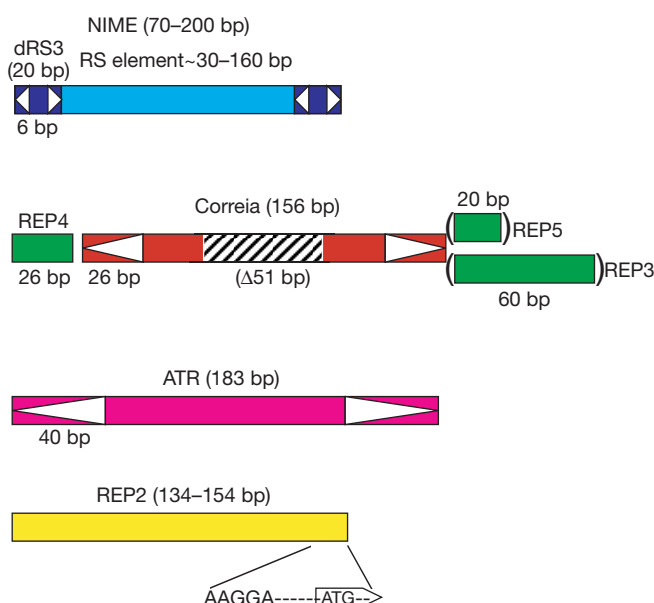


Figure 2 Types of *N. meningitidis* repeat. The name of each repeat type is indicated above the repeat. Inverted repeat sequences are represented by open triangles, and the internal deletion present in some Correia elements by a hatched box. The translational signals within REP2 are indicated below the repeat.

Table 1 Number of repeats by type in *N. meningitidis* Z2491

Type	Size (bp)	Frequency
DNA uptake sequence: gccgtctgaa	10	1,892
RS	24–161	681
dRS3: attccnnnnnnnnnggaat	20	772
Correia (full)	150–159	173
Correia (internal deletion)	~104	84
Correia (partial)	37–145	29
ATR	183	19
REP 2	59–154	26
REP 3	60	13
REP 4	26	20
REP 5	20	9
IS1016	256–740	14 (including partial)
IS1106	263–1219	22 (including partial)
IS1655	1,074–1,257	7 (including partial)
Prophage	2,330–38,964	5

products of these genes are potential targets for the immune system, and antigenic variation of these proteins would be beneficial in evading the immune response. The genes encoding the outer membrane proteins PorA (NMA1642) and PorB (NMA0398) are flanked on both sides by arrays of repeats, as are *lbpAB* (ref. 13) (Fig. 3b), *tbpAB* (ref. 14) and *hpuAB* (ref. 15), which encode lactoferrin-, transferrin- and haemoglobin/haptoglobin-binding proteins, respectively. The putative ferric enterobactin-binding protein genes *fetABI* are also associated with two such arrays. In contrast, the neisserial ferric binding protein operon *fbpABC* (ref. 16), which may not be exposed to the immune response, the TonB-dependent receptor NMA0577 and the haemoglobin receptor *hmbR* (ref. 17) are not associated with repeat elements (a poly-(G) tract is present in *hmbR*, potentially allowing phase-variable expression; see below). The *opaD*, *opaB*, *opaA* and *opcA* genes, which are involved in various interactions with the host cell, are all flanked on one or both sides by repeat arrays. Comparison of these regions

from a range of *N. meningitidis* isolates has shown that the repeat arrays themselves are highly variable between diverse *Neisseriae* but may be identical in related strains even after more than 30 years of epidemic spread (G.M., unpublished data).

Repeat-mediated gene rearrangements are also apparent at other sites, although, in these cases, the repeats are local. One such site encompasses the *mafA* and *mafB* genes; *mafA* is predicted to be a lipoprotein and *mafB* to be secreted; both have been proposed to have adhesin activity in *N. gonorrhoeae* (S. Eickernjaeger *et al.*, personal communication; EMBL accession number AF142582). The *N. meningitidis* orthologue of *mafB* is immediately followed by a 6.5-kb region of low G+C DNA encoding mainly short genes of unknown function. Interspersed with these genes, and with a normal G+C content, are three repeats of 303 bp and one of 79 bp from *mafB*, each corresponding to the beginning of an open reading frame (ORF) that has no start codon. This arrangement suggests that these repeat sequences may be capable of recombination

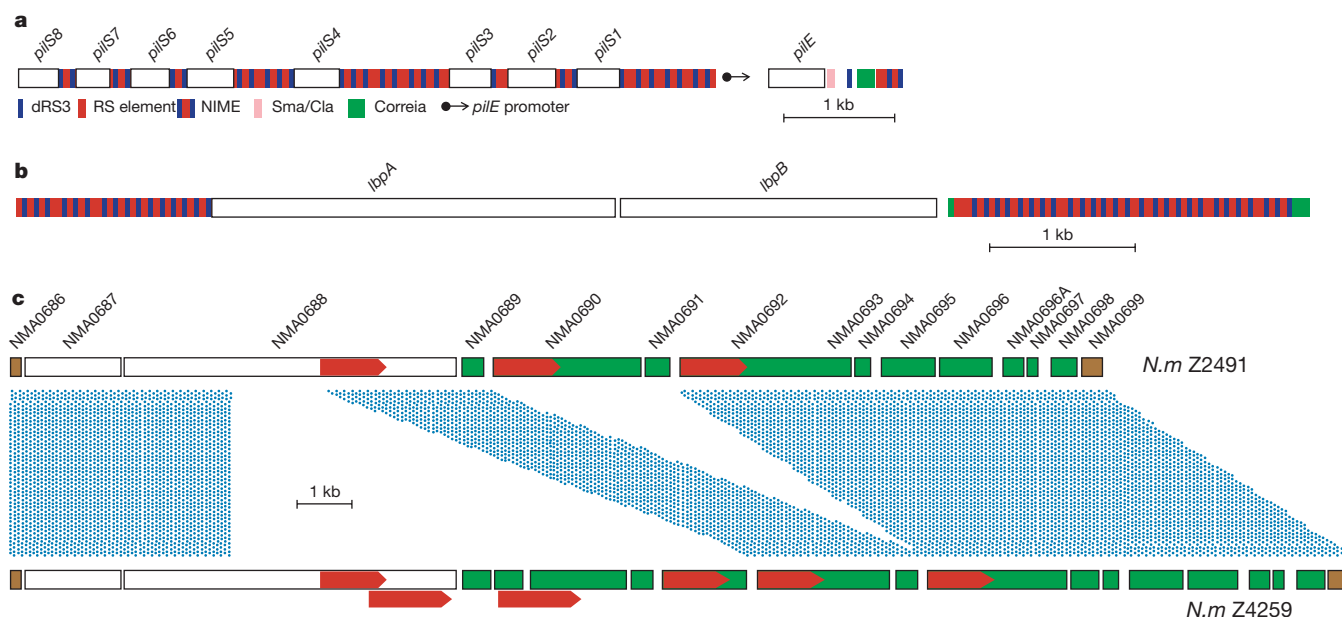


Figure 3 Structure of selected repeat regions. **a**, Repeats around the *pilE/S* locus. *pil* genes are indicated by open boxes and repeats by coloured boxes as described in the key. **b**, Large repeat arrays around the *lbpA* and *lbpB* genes. Repeats are coloured as in **a**. **c**, The filamentous haemagglutinin homologue gene sequences of *N. meningitidis* Z2491 and Z4259. Conserved sequences are indicated by light blue bars, and internal repeat

regions by red arrows. The complex repeat structures have been simplified for clarity. Open boxes represent genes encoding surface-exposed proteins, and green boxes represent genes with no database matches. The brown boxes represent the two halves of an ABC transporter pseudogene.

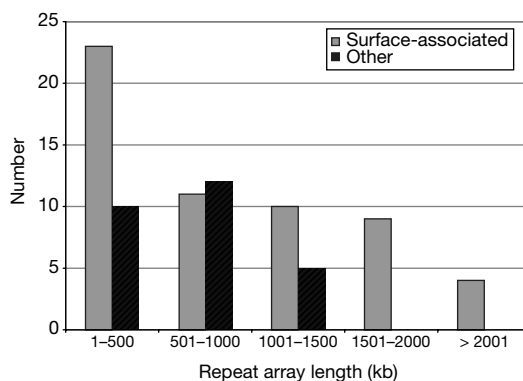


Figure 4 Graph showing the relationship between repeat array length and flanking gene function. For each category of repeat array length the number of flanking genes

associated with surface structures is shown in grey, and the number of genes in all other categories is shown hatched.

with the *mafB* gene, leading to the attachment of different 3' sequences. This interpretation is supported by the observation that the 3' sequence of the gonococcal *mafB* orthologue is almost identical to the ORF following the first 303-bp repeat, and not to the 3' sequence of the meningococcal *mafB* gene.

A second example of repeat-mediated rearrangement of the 3' ends of genes encoding surface-exposed proteins is provided by NMA0688, a homologue¹⁸ of the *Bordetella pertussis* filamentous haemagglutinin *fhaB* (ref. 19). The sequence downstream of NMA0688 contains complex repeated sequences from NMA0688 (Fig. 3c) and encodes several proteins of unknown function. Sequencing this region from *N. meningitidis* Z4259 (a serogroup C strain; EMBL accession number AJ391284) revealed a similar repeat structure, but with additional DNA inserted into the 3' sequence of NMA0688 at one of the repeats. This has the effect of changing the 3' end of the NMA0688 gene, as well as introducing additional genes, some containing similar repeats. The 3' region of the Z4291 NMA0688 gene is within a downstream ORF in Z4259. It therefore seems likely that, like *mafB*, the 3' end of this *fhaB* homologue can be altered by recombination with alternative 3' ends. Intriguingly, the entire region consisting of the *fhaC* (NMA0687) and *fhaB* homologues, along with the downstream DNA, appears to have been inserted into the coding sequence of an ABC transporter pseudogene (NMA0686/NMA0699).

Tandem repeats of nucleotides in *Neisseria* are subject to lengthening or shortening during replication owing to a phenomenon called slipped-strand mispairing²⁰. When these tandem repeats are present in coding sequences, this change in length can change the translation state of the gene, leading to an on/off switching of the gene product called 'phase variation'. The complete genome sequence reveals around 26 tandem repeats indicating potentially phase-variable genes (Table 3, Supplementary Information), many of which have been studied before. These repeats range from homopolymeric tracts of G or C nucleotides to di-, tetra- and penta-nucleotide repeats. As would be expected, most phase-variable genes are either surface-exposed or involved in the biosynthesis or modification of surface structures, which supports their suggested role in virulence and/or immune avoidance. Four of the tandem repeats are not directly within coding sequences, but are immediately upstream of candidate variable genes, and these may affect transcription from gene-specific promoters, as has been shown for *opcA* (ref. 21) and *porA* (ref. 22).

The analysis presented here supports the general operation of three mechanisms of repeat-mediated antigenic variation within the *N. meningitidis* genome: on/off switching and transcriptional modulation of gene expression by slipped-strand mispairing of short tandem repeats; intragenomic recombination of localized repeats leading to the use of different carboxy termini for surface-exposed proteins; and intergenomic gene conversion of specific surface-associated genes associated with large arrays of global repeats, mediated by the internalization of related DNA through the highly repetitive DNA uptake sequence. □

Methods

DNA was prepared from *N. meningitidis* Z4291 as described²³. The DNA was fragmented by sonication, size-fractionated on an agarose gel, and two libraries were generated in pUC18 using size fractions ranging from 0.5 to 1.5 kb. Roughly 37,500 pUC clones were sequenced from both ends using Dye-terminator chemistry on ABI 373 and 377 sequencing machines; 58,269 reads were used to generate the final assembly, giving about a 10-fold coverage of the genome. Sequence assembly was accomplished using Phrap (P. Green, unpublished), and the sequencing was finished using GAP4 (ref. 24). The assembly was verified by genomic PCR reactions across all unbridged large repeats, in addition to 260 forward and reverse reads from a random library of 20–22 kb cloned in lambdaFixII (Stratagene), and 610 forward and reverse reads from 3 libraries of 9–13 kb cloned in pSP64 (Promega). The final assembly was checked against the published map using the positions of restriction sites, mapped genes and lambda clones. In the final assembly, less than 0.07% of the genome was covered by a single clone only, and less than 0.12% was unsequenced on both strands, or with complementary sequencing chemistries. The DNA was compared with sequences in the EMBL database using BLASTN and

BLASTX²⁵. Transfer RNAs were predicted by tRNAscan-SE (ref. 26). Potential CDSs were predicted using ORPHEUS²⁷ and GLIMMER²⁸ (both trained on an initial ORF set generated by ORPHEUS), and the results were combined and checked manually. The predicted protein sequences were searched against a non-redundant protein database using WUBLASTP and FASTA. The complete six-frame translation was used to search PROSITE, and the predicted proteins were compared against the PFAM database of protein domain hidden Markov models²⁹. The results of all these analyses were assembled together using the Artemis sequence viewer (K.M.R., unpublished) and used to inform a manual annotation of the sequence and predicted proteins. Annotation was based, wherever possible, on characterized proteins or genes. Repeat sequences were identified using HMMER (S. Eddy, personal communication) and the EMBOSS program 'profit' (<http://www.sanger.ac.uk/Software/EMBOSS/>), based on sequence alignments generated with ClustalW³⁰.

Received 1 March; accepted 8 March 2000.

- Dempsey, J. A., Wallace, A. B. & Cannon, J. G. The physical map of the chromosome of a serogroup A strain of *Neisseria meningitidis* shows complex rearrangements relative to the chromosomes of the two mapped strains of the closely related species *N. gonorrhoeae*. *J. Bacteriol.* **177**, 6390–6400 (1995).
- Curnow, A. W., Tumbula, D. L., Pelaschier, J. T., Min, B. & Soll, D. Glutamyl-tRNA (Gln) amidotransferase in *Deinococcus radiodurans* may be confined to asparagine biosynthesis. *Proc. Natl Acad. Sci. USA* **95**, 12838–12843 (1998).
- Andersson, S. G. et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
- Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665 (1996).
- Smith, N. H., Holmes, E. C., Donovan, G. M., Carpenter, G. A. & Spratt, B. G. Networks and groups within the genus *Neisseria*: analysis of *argF*, *recA*, *rho*, and 16S rRNA sequences from human *Neisseria* species. *Mol. Biol. Evol.* **16**, 773–783 (1999).
- Swartley, J. S. et al. Characterization of the gene cassette required for biosynthesis of the (α1–6)-linked N-acetyl-D-mannosamine-1-phosphate capsule of serogroup A *Neisseria meningitidis*. *J. Bacteriol.* **180**, 1533–1539 (1998).
- Goodman, S. D. & Socca, J. J. Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc. Natl Acad. Sci. USA* **85**, 6982–6986 (1988).
- Correia, F. F., Inouye, S. & Inouye, M. A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae*. *J. Biol. Chem.* **263**, 12194–12198 (1988).
- Haas, R. & Meyer, T. F. The repertoire of silent pilos genes in *Neisseria gonorrhoeae*: Evidence for gene conversion. *Cell* **44**, 107–115 (1986).
- Mehr, I. J. & Seifert, H. S. Differential roles of homologous recombination pathways in *Neisseria gonorrhoeae* pilin antigenic variation, DNA transformation and DNA repair. *Mol. Microbiol.* **30**, 697–710 (1998).
- Howell-Adams, B., Wainwright, L. A. & Seifert, H. S. The size and position of heterologous insertions in a silent locus differentially affect pilin recombination in *Neisseria gonorrhoeae*. *Mol. Microbiol.* **22**, 509–522 (1996).
- Wainwright, L. A., Pritchard, K. H. & Seifert, H. S. A conserved DNA sequence is required for efficient gonococcal pilin antigenic variation. *Mol. Microbiol.* **13**, 75–87 (1994).
- Pettersson, A., Prinz, T., Umar, A., van der Biezen, J. & Tommassen, J. Molecular characterization of LbpB, the second lactoferrin-binding protein of *Neisseria meningitidis*. *Mol. Microbiol.* **27**, 599–610 (1998).
- Vonder Haar, R. A., Legrain, M., Kolbe, H. V. & Jacobs, E. Characterization of a highly structured domain in Tbp2 from *Neisseria meningitidis* involved in binding to human transferrin. *J. Bacteriol.* **176**, 6207–6213 (1994).
- Lewis, L. A., Gray, E., Wang, Y. P., Roe, B. A., & Dyer, D. W. Molecular characterization of hpuAB, the haemoglobin-haptoglobin-utilization operon of *Neisseria meningitidis*. *Mol. Microbiol.* **23**, 737–749 (1997).
- Adhikari, P., Berish, S. A., Nowalk, A. J., Veraldi, K. L., Morse, S. A. & Mietzner, T. A. The *fhpABC* locus of *Neisseria gonorrhoeae* functions in the periplasm-to-cytosol transport of iron. *J. Bacteriol.* **178**, 2145–2149 (1996).
- Stojiljkovic, I. et al. The *Neisseria meningitidis* haemoglobin receptor: its role in iron utilization and virulence. *Mol. Microbiol.* **15**, 531–541 (1995).
- Klee, S. R. et al. Molecular and biological analysis of eight genetic islands that distinguish *Neisseria meningitidis* from the closely related pathogen *Neisseria gonorrhoeae*. *Infect. Immunity* (in the press).
- Domenighini, M. et al. Genetic characterization of *Bordetella pertussis* filamentous haemagglutinin: a protein processed from an unusually large precursor. *Mol. Microbiol.* **4**, 787–800 (1990).
- Henderson, I. R., Owen, P. & Nataro, J. P. Molecular switches—the ON and OFF of bacterial phase variation. *Mol. Microbiol.* **33**, 919–932 (1999).
- Sarkari, J., Pandit, N., Moxon, E. R. & Achtman, M. Variable expression of the Opc outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing poly-cytidine. *Mol. Microbiol.* **13**, 207–217 (1994).
- van der Ende, A. et al. Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the spacing between the -10 and -35 regions of the promoter. *J. Bacteriol.* **177**, 2475–2480 (1995).
- Zhou, J. & Spratt, B. G. Sequence diversity within the *argF*, *fhp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Mol. Microbiol.* **6**, 2135–2146 (1992).
- Bonfield, J. K., Smith, K. F. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
- Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26**, 2941–2947 (1998).
- Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26**, 544–548 (1998).

29. Bateman, A. *et al.* Pfam 3.1: 1,313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27**, 260–262 (1999).
30. Higgins, D. G., Bleasby, A. J. & Fuchs, R. CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**, 189–191 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This research was funded by The Wellcome Trust.

Correspondence and requests for materials should be addressed to J.P. (e-mail: parkhill@sanger.ac.uk). The complete sequence and annotation can be obtained from the EMBL database with the ID NMAZ2491 (accession number AL157959), and from our web pages (http://www.sanger.ac.uk/Projects/N_meningitidis).

The duration of antigen receptor signalling determines CD4⁺ versus CD8⁺ T-cell lineage fate

Koji Yasutomo*, Carolyn Doyle†, Lucio Miele‡ & Ronald N. Germain*

* Lymphocyte Biology Section, Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

† Department of Immunology, Duke University Medical Centre, Durham, North Carolina 27710, USA

‡ Cancer Immunology Program, Cardinal Bernardin Cancer Centre, Loyola University Medical Centre, Maywood, Illinois 60153, USA

Signals elicited by binding of the T-cell antigen receptor and the CD4/CD8 co-receptor to major histocompatibility complex (MHC) molecules control the generation of CD4⁺ (helper) or CD8⁺ (cytotoxic) T cells from thymic precursors that initially express both co-receptor proteins¹. These precursors have unique, clonally distributed T-cell receptors with unpredictable specificity for the self-MHC molecules involved in this differentiation process². However, the mature T cells that emerge express only the CD4 (MHC class II-binding) or CD8 (MHC class I-binding) co-receptor that complements the MHC class-specificity of the T-cell receptor. How this matching of co-receptor-defined lineage and T-cell-receptor specificity is achieved remains unknown^{1,3,4}, as does whether signalling by the T-cell receptors, co-receptors and/or general cell-fate regulators such as Notch-1 (refs 5, 6) contributes to initial lineage choice, to subsequent differentiation processes or to both. Here we show that the CD4 versus CD8 lineage fate of immature thymocytes is controlled by the co-receptor-influenced duration of initial T-cell receptor-dependent signalling. Notch-1 does not appear to be essential for this fate determination, but it is selectively required for CD8⁺ T-cell maturation after commitment directed by T-cell receptors. This indicates that the signals constraining CD4 versus CD8 lineage decisions are distinct from those that support subsequent differentiation events such as silencing of co-receptor loci.

The AND T-cell receptor (TCR) is specific for a pigeon cytochrome *c* peptide bound to the MHC class II molecule I-E^k (ref. 7). Transgenic thymocytes expressing this TCR are efficiently selected into the CD4⁺ lineage in mice expressing wild-type I-A^b MHC class II molecules. In contrast, AND TCR transgenic mice expressing mutant I-A^b molecules that are defective in interaction with CD4 but not the TCR⁸ generate CD8⁺ but not CD4⁺ mature cells⁹. To investigate how altering CD4 co-receptor binding controls the lineage fate of AND thymocytes, we took advantage of a modified two-stage reaggregate culture system¹⁰ that allows controlled deliv-

ery of MHC-dependent and independent signals to thymocytes at distinct stages of maturation (Yasutomo *et al.*, manuscript in preparation). Immature CD69^{lo}CD4⁺CD8⁺ TCR transgenic thymocytes (double positive; DP) are incubated in dispersed culture with cells expressing the desired MHC molecule ligands, to initiate selection events (culture 1). This first TCR stimulation does not lead to silencing of the CD4 or CD8 locus in dispersed culture for up to three days, but it does rapidly induce upregulation of the activation marker CD69 that characterizes thymocytes undergoing maturation *in vivo*^{1,11}. When the CD69^{med/hi} thymocytes arising in culture 1 after 20 h are purified and reaggregated with either MHC-positive thymic stromal cells (TSC) or MHC-negative TSC in the presence of dendritic cells of the appropriate MHC type (culture 2), mature functional T cells expressing only a single co-receptor develop over the next 60 h. Experiments using this model showed that TCR–MHC molecule interactions are required not only in culture 1 to generate the CD69^{hi} thymocytes, but also in culture 2 to generate mature T cells. Using this approach, we asked whether the first, the second or both sets of TCR/co-receptor–MHC interaction events determined the differentiation of AND TCR transgenic T cells along the CD4 versus CD8 pathways.

Thymocytes from mice expressing only the AND TCR in the absence of MHC class II molecules (AND TCR transgenic RAG-2^{−/−} MHC Aβ^{−/−}, referred to as AND throughout) were stimulated in culture 1 using dendritic cells from wild-type mice (WT-DC) or mutant (Mu-DC) I-A^b transgenic mice. When purified cells of the CD69⁺CD4⁺CD8⁺ phenotype (CD69⁺DP) generated by this stimulation were reaggregated with TSC from MHC class II^{−/−} mice, no CD4⁺ or CD8⁺ mature T cells developed (Fig. 1a, b). Inclusion of

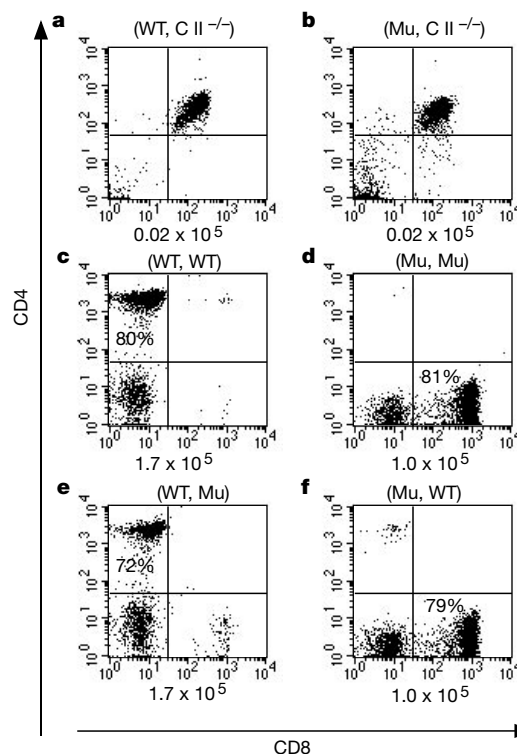


Figure 1 Role of co-receptor in CD4/CD8 lineage choice. Sorted CD69⁺CD4⁺CD8⁺ cells obtained after stimulation by wild-type dendritic cells (WT-DC; **a, c, e**) or mutant dendritic cells (Mu-DC; **b, d, f**) were cultured with MHC class II^{−/−} TSC (**a, b**) or MHC^{−/−} TSC and wild-type dendritic cells (WT; **c, f**), or mutant dendritic cells (Mu; **d, e**) for 72 h in thymic reagggregates. The recovered cells from all cultures were stained with anti-CD4 or CD8 monoclonal antibodies and examined by flow cytometry. The percentage of cells with a CD4⁺ or CD8⁺ phenotype is given in the upper left or lower right quadrant of each panel and the absolute number of recovered viable thymocytes is given below each panel.

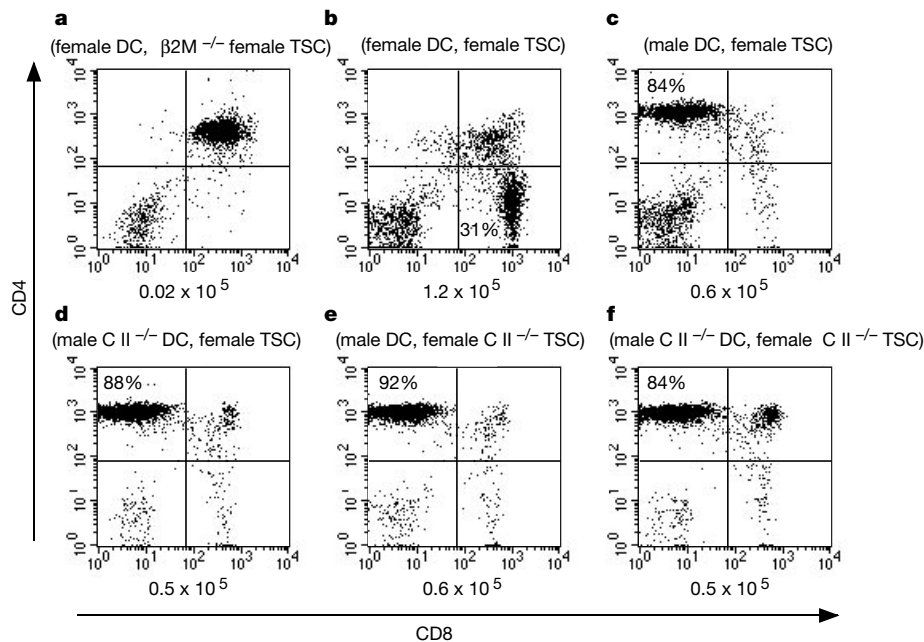


Figure 2 Potent TCR ligands can promote thymocytes with an MHC class-I-restricted TCR to become CD4⁺ T cells. DP thymocytes from female HY TCR transgenic H-2^d RAG-2^{-/-} mice were stimulated for 20 h with splenic dendritic cells (DC) from female (a, b) or male (c, e) C57BL/6 mice or splenic DC from male C57BL/6 crossed with MHC class II^{-/-} mice (d, f) (culture 1). The CD69^{hi} cells generated in culture 1 were placed in reaggregate culture for 72 h with TSC from female $\beta 2M^{-/-}$ (a), female C57BL/6 WT (b-d) or female MHC class II^{-/-} (e, f) mice (culture 2). Cells were analysed as in Fig. 1.

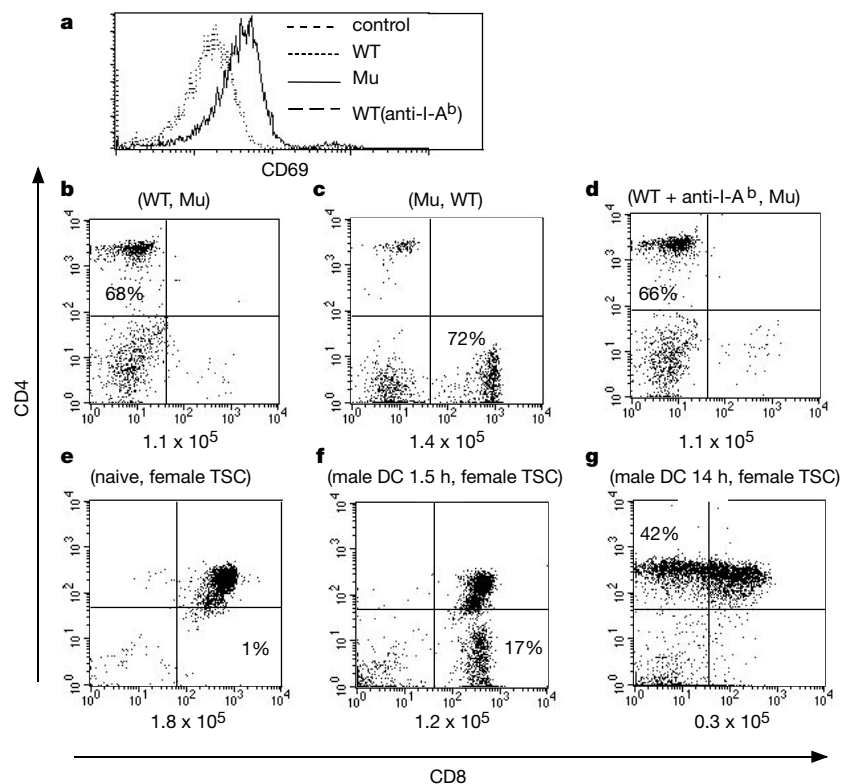


Figure 3 The duration, not the magnitude, of TCR signalling dictates lineage choice. a, DP thymocytes from AND mice were stimulated for 20 h with Mu-DC or WT-DC in the absence or presence of anti-I-A^b monoclonal antibodies (3JP). Purified CD69^{hi}CD4^{lo}CD8^{lo} cells after stimulation by Mu-DC (c), or WT-DC in the absence (b) or presence (d) of anti-I-A^b monoclonal antibodies, were then cultured with TSC from MHC class II^{-/-} mice and WT-DC (e) or Mu-DC (b, d) for 60 h. DP thymocytes from female HY TCR transgenic H-2^d RAG-2^{-/-} mice were stimulated with splenic dendritic cells from male C57BL/6 mice for 0 (e), 1.5 (f) or 14 (g) hours. The CD69^{hi} cells generated by culture were placed in reaggregate culture with TSC from female C57BL/6 mice for 60 (e), 58 (f) or 46 (g) hours. Cells were analysed as in Fig. 1.

either WT-DC or Mu-DC in culture 2 yielded mature T cells (Fig. 1c–f). However, the CD69⁺DP generated using WT-DC differentiated exclusively into CD4⁺ cells, whereas CD69⁺DP produced using Mu-DC developed only into CD8⁺ cells, irrespective of which dendritic cells were used in culture 2.

These findings are inconsistent with previously described stochastic/selection models for CD4/8 lineage commitment and mature T-cell development^{4,12}. Such models propose that co-receptor and TCR co-recognition of MHC molecules is required during late stages of development to maintain cell viability, but our results show that CD4 function is not required for CD4⁺ T-cell development in culture 2 (Fig. 1e). We do, however, find that MHC class I–CD8 interactions are required for generation of mature CD8⁺ cells in culture 2 (data not shown). In addition, both wild-type and mutant MHC class II molecules can support either CD4⁺ or CD8⁺ T-cell development in culture 2 (Fig. 1d, f). If the initial TCR signal produced a stochastic pattern of commitment, then, unless stringent selection for survival takes place in culture 1, both CD4⁺ and CD8⁺ T-cell differentiation should have been seen in culture 2. This was not the case (Fig. 1e, f). Finally, selection during culture 1 based on inadequate signalling resulting from mismatches between co-receptor expression and TCR binding specificities for MHC molecules cannot explain these findings, because neither CD4 nor CD8 expression is extinguished at this time¹¹ (Yasutomo *et al.*, manuscript in preparation).

DP cells have a higher fraction of CD4 as compared to CD8 linked to the key src-family kinase Lck, which is involved in TCR-dependent signalling^{13,14}. This asymmetry could explain why most cells with TCR that bind MHC class II ligands co-recognized by CD4 were restricted to the CD4 pathway, whereas those with TCR engaging MHC class I ligands along with CD8 were constrained to the CD8 lineage. To test this presumption, we investigated whether exposure to a potent TCR agonist that efficiently recruits

CD8 (ref. 15) could overcome this biased Lck distribution and generate CD69⁺DP expressing MHC class I-restricted TCR that would develop along the CD4 pathway. The deletion that accompanies continuous exposure to such strong ligands in typical thymic organ cultures¹⁶ can be avoided in our model by excluding agonist ligands from culture 2 (Yasutomo *et al.*, manuscript in preparation).

HY-TCR transgenic thymocytes are deleted in male mice, owing to expression of the agonist HY male antigen presented by H-2D^b, whereas they are positively selected into the CD8 lineage in female H-2^b mice¹⁷. When CD69⁺DP generated in culture 1 by the stimulation of HY-TCR transgenic thymocytes with female H-2^b dendritic cells were reaggregated in culture 2 with female H-2^b TSC (Fig. 2b), only CD8⁺ T cells were produced. Like AND thymocytes stimulated with Mu-DC, CD8⁺ T cells did not develop from these CD69⁺DP when TSC from female $\beta 2m$ -deficient mice were used (Fig. 2a). Remarkably, when we placed CD69^{hi} HY-TCR transgenic cells generated by stimulation with male H-2^b dendritic cells in culture 2 with TSC from female C57BL/6 mice, CD4⁺ and not CD8⁺ T cells developed (Fig. 2c). This development of CD4⁺ T cells from HY-TCR transgenic thymocytes did not require recognition of MHC class II molecules in either culture 1 or 2 (Fig. 2d–f). Maturation into CD4⁺ T cells can be completed even though CD8 expression is lost during this process, providing additional evidence against selection schemes postulating that survival of maturing thymocytes demands continuous signalling from ligand co-recognition by TCR and co-receptor.

Published models postulating that DP lineage decisions are controlled by a unique biochemical signal from Lck-associated co-receptors did not predict this generation of CD4⁺ T cells from MHC class I-restricted TCR transgenic thymocytes using a potent agonist¹⁸, re-opening the question of how co-receptor function influences lineage fate. We first investigated whether a change in ligand density, and hence in peak signalling intensity, controls

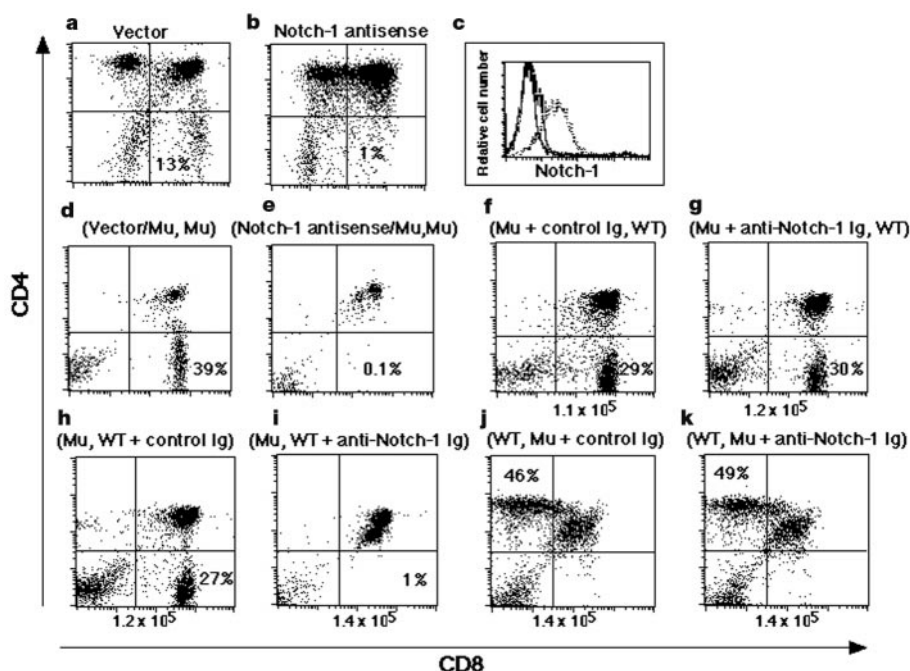


Figure 4 Notch-1 is involved in post-commitment maturation of CD8 lineage T cells. Day 14 fetal thymocytes from C57BL/6 mice (a–c) or AND mice (d, e) were infected with retrovirus encoding GFP (a, c, d) or GFP and anti-sense mouse Notch-1 (64–1164) (b, e) and cultured as described in the Methods. Cells infected by vector alone or vector encoding antisense mouse Notch-1 were stained with a combination of anti-Notch-1, anti-CD4 and anti-CD8 monoclonal antibodies (c) or by anti-CD4 and anti-CD8 monoclonal antibodies (a, b, d, e). Notch-1 expression in (c) is gated on CD4⁺CD8⁺ cells

(sparse dot, vector; dense dot, antisense Notch-1; solid, control). The cells in a, b, d, e are gated on GFP⁺ cells. DP thymocytes from AND mice were also stimulated for 20 h with Mu-DC in the presence of mouse IgG2b (f) or anti Notch-1 mAb (g), Mu-DC (h, i), or WT-DC (j, k). The sorted CD69⁺CD4⁺CD8⁺ cells obtained after stimulation were cultured for 60 h with MHC^{−/−} TSC and WT-DC (f, g), WT-DC in the presence of mouse IgG2b (h) or anti-Notch-1 antibodies (i), or Mu-DC in the presence of mouse IgG2b (j) or anti-Notch-1 antibodies (k).

lineage fixation. The interaction between the AND TCR and I-A^b was blocked with an anti-MHC class II monoclonal antibody, using a concentration of antibody that limits CD69 expression induced by WT-DC to the same level as that induced by Mu-DC without antibody (Fig. 3a). The thymocytes thus serve as their own control that this is an appropriate level of blocking, because the level of CD69 upregulation on DP has been shown to track the density of offered ligand and hence to reflect TCR occupancy¹⁹ (K. Yasutomo, unpublished observations). Cells exposed to less wild-type signal with CD4 function intact still entered the CD4 lineage (Fig. 3d). These data indicate that the CD4 co-receptor does more than simply increase the absolute level of MHC-dependent signalling.

A previous study using pharmacologic agents indicated that the duration of active signal generation influenced the CD4/CD8 decision process²⁰. We therefore examined the effects of varying the length of time for which the TCR could engage ligand. When CD69⁺ cells generated by the stimulation of HY TCR transgenic thymocytes using male H-2^b dendritic cells for 1.5 h were placed in culture 2 with female H-2^b TSC, only CD8⁺ T cells developed (Fig. 3f). Under the same culture conditions, CD8⁺ T cells do not mature from naive DP thymocytes (Fig. 3e). In contrast, after 14 h of initial stimulation, only CD4⁺ T cells appeared (Fig. 3g). This indicates that effective co-receptor activity may constrain DP development to the CD4 lineage by helping to ensure signalling over an adequate time interval, consistent with recent observations on the role of co-receptor activity in forming an 'immunological synapse' with an antigen-presenting cell that is necessary for sustained TCR signalling²¹.

Signals from receptors other than those recognizing MHC molecules have been proposed to influence CD4/CD8 commitment. Notch is central in cell-fate determination in many tissues and some data indicate that activated Notch-1 may promote CD8 lineage choice². Others have suggested that Notch-1 signalling simply prolongs survival of all TCR-signalled DP thymocytes⁶. Gene targeting of Notch-1 interferes with T-cell development before the DP stage, so such mutant animals or their cells cannot be used to resolve this issue²². The preceding evidence showing a distinction between lineage commitment and lineage progression allowed us to re-examine this issue. First, we used a retrovirus encoding Notch-1 in the antisense orientation to infect fetal thymocytes. The infected cells produced normal numbers of CD4⁺ T cells in organ culture but CD8⁺ T cell development was severely impaired (Fig. 4a, b). The antisense vector was highly effective, reducing Notch-1 expression on the emerging CD4⁺ cells by more than 90% (Fig. 4c). Likewise, when the same retrovirus was used to infect AND thymocytes and these cells were stimulated by Mu-DC, very few DP became CD8⁺ T cells, with no compensatory appearance of CD4⁺ cells (Fig. 4d, e).

The second set of experiments employed an anti-Notch-1 monoclonal antibody. When AND CD69⁺DP generated by culture 1 stimulation with Mu-DC were reaggregated in culture 2 with MHC class II^{-/-} TSC and WT-DC, CD8⁺ T cells developed (Fig. 4f). If anti-Notch-1 monoclonal antibodies were added to culture 1, thymocytes still became CD8⁺ in similar numbers (Fig. 4g). However, when anti-Notch-1 monoclonal antibodies were added to culture 2, differentiation to CD8⁺ SP (single positive) cells was very limited (Fig. 4h, i). Anti-Notch monoclonal antibodies did not interfere with the development of CD4⁺ T cells when present at either culture step (Fig. 4j, k). Together, the antisense and antibody blocking data show that Notch-1 activity is selectively involved in CD8 but not CD4 lineage progression after lineage fate is fixed by TCR/co-receptor signalling.

These results indicate that the duration of initial TCR signalling determined by co-receptor-influenced self-MHC recognition rapidly fixes thymocyte lineage fate; lineage choice does not appear to be stochastic or cell autonomous. The simplest model that fits our data is an instructive one, with signal duration controlling which set of genes is activated or silenced so that a DP

thymocyte adopts either of two possible fates. Such a model is consistent with data showing that signal duration can influence gene expression and cellular differentiation in non-lymphoid systems²³. Alternatively, DP could consist of two subpopulations, each committed to either the CD4 or CD8 lineage. Long-lasting signals would be necessary to induce the CD4 program and, to fit our data, would also need to kill pre-CD8 cells. Short signals would activate only the pre-CD8 cells, without affecting pre-CD4 thymocytes. Without direct evidence for bipotency of immature thymocytes, this alternative cannot be dismissed.

In either case, these observations provide a specific mechanism for previous results showing that a 'strength of signal' parameter affects CD4/CD8 choice²⁴ and that antibody crosslinking of CD3 ϵ with either CD4 or CD8 drives CD4 lineage development, whereas CD3 ϵ crosslinking alone produces CD8⁺ T cells¹⁸. The data also connect our demonstration that mitogen-activated protein kinase (MAPK)-mediated positive feedback enhances the duration of TCR signalling (Stefanova *et al.*, manuscript in preparation) to evidence that interference with the MAPK pathway promotes CD8 lineage choice^{25,26}. The lack of a unique fate-determining biochemical signal arising from CD4 versus CD8 engagement implies that errors will be made in matching TCR class-specificity with lineage choice. Errors in the direction of CD4 commitment will probably involve very potent MHC class I ligands for the TCR that will induce late-stage deletion. Errors in the converse direction should be less frequent and cells with MHC class II-specific TCR that traverse the CD8-dependent progression steps are in any case unlikely to be dangerous to the host.

A final finding of our study is the clear distinction between lineage commitment itself and subsequent lineage-specific differentiation events such as selective silencing of co-receptor expression. This new insight allowed us to show that Notch-1 is critical only for supporting CD8 lineage progression after lineage fate is fixed. Because several Notch family members are expressed by T-cell precursors and both they and their putative ligands show complex patterns of regulation²⁷ and functional interactions²⁸, other Notch receptors may have a complementary function in the CD4 pathway to that shown here for Notch-1 in CD8 differentiation. The methods we have described provide tools for direct analysis of this possibility. □

Methods

Mice

Mice with targeted inactivation of both the $\beta 2$ microglobulin gene and/or A β^b gene loci and HY TCR transgenic RAG-2^{-/-} mice were obtained from Taconic. B10.BR and C57BL/10 mice were obtained from the Jackson Laboratories. The AND TCR transgenic⁷ RAG-2^{-/-} A β^b ^{-/-} mice were provided by B. J. Fowlkes and are referred to as AND throughout. Transgenic mice expressing a mutant I-A β^b chain with the E137A and V142A mutations that interfere with CD4 binding (mutant MHC class II) or expressing the corresponding WT I-A β^b chain, each bred to A β^b ^{-/-} mice, have been described⁹. All mice were bred and maintained in accordance with established guidelines.

Flow cytometry

Thymocytes were stained with PE-conjugated anti-CD4 and FITC-conjugated anti-CD8 monoclonal antibodies. All antibodies were purchased from Pharmingen. Flow cytometry was performed on a FACScan or a FACStarPlus (Becton Dickinson). List-mode data files were analysed using Cell Quest software (Becton Dickinson).

Thymocyte stimulation and preparation of CD69^{hi} cells

CD4⁺CD8⁺ thymocytes (1×10^6) from 1–2 week old AND mice or HY TCR transgenic RAG-2^{-/-} nonselecting B10.D2 mice were mixed with dendritic cells (2×10^5) in the absence of added antigen, then cultured for 1.5 or 20 h at 37 °C. Viable CD69^{hi} cells for transfer into reaggregate cultures were obtained by fluorescence-activated cell sorting after stimulation of CD4⁺CD8⁺ thymocytes using the described culture conditions. In some experiments purified mouse IgG2b ($20 \mu\text{g ml}^{-1}$) (Pharmingen), 3Jp monoclonal antibodies against I-A^b ($0.12 \mu\text{g ml}^{-1}$), or anti-Notch-1 monoclonal antibodies ($20 \mu\text{g ml}^{-1}$) (L.M., manuscript in preparation) were added during stimulation.

Thymic reaggregate culture

Thymic stromal cells (TSC) were prepared by disaggregating fetal thymic lobes previously

cultured for 5 days in 1.35 mM deoxyguanosine (Sigma) using 0.05% trypsin (GIBCO) and 0.02% EDTA. Reaggregates were formed by mixing together the desired TSC and thymocytes at a 1:1 cell ratio (absolute number, 5×10^5 of each cell type) or TSC, thymocytes and dendritic cells at a cell ratio of 10:10:1 (absolute number of dendritic cells, 5×10^4). After pelleting the cells by centrifugation, the cell mixture was placed as a standing drop on the upper membrane surface of a Transwell culture well containing RPMI-1640 supplemented by 10% FCS and cultured for 72 h at 37 °C. In some experiments purified mouse IgG2b (20 $\mu\text{g ml}^{-1}$) (Pharmingen) or anti-Notch-1 monoclonal antibodies (20 $\mu\text{g ml}^{-1}$) was incubated with CD69^{hi} thymocytes, TSC and dendritic cells for 30 min at 4 °C before reaggregate culture, then cell mixtures were cultured in the presence of the same antibody. All results shown were repeated 2–8 times with similar findings.

Preparation of dendritic cells

Dendritic cells were purified from spleen cells from C57BL/6 or B10.BR mice as described²⁹. Briefly, low-density spleen cells recovered from a BSA gradient were incubated for 30 min with anti-CD4, anti-CD8 and anti-B220 monoclonal antibodies (Pharmingen). After incubation with this antibody cocktail, the cells were washed and antibody-coated cells were removed by sheep-anti-rat IgG-coupled magnetic beads (Dynabeads, Dynal). The resultant population contained 80–85% CD11c⁺ cells and is referred to as dendritic cells (DC) in this study.

Retroviral constructs and retroviral transfection

The retroviral vector encoding GFP (GFP–RV)³⁰ was provided by K. Murphy (Washington Univ., St. Louis). Mouse Notch-1 (+64–+1164) was amplified using Pfu DNA polymerase (Stratagene) and cloned into *XhoI* digested GFP–RV vector. The nucleotide sequence and orientation of the insert were confirmed by dideoxy sequencing. The Phoenix-Eco packaging cell line (gift from G. Nolan, Stanford) was transfected according to Nolan's protocol. Day 14 fetal thymocytes ($5 \times 10^6 \text{ ml}^{-1}$) from C57BL/6 or AND mice were infected using 1:1 volume of viral supernatant, polybrene (Sigma) at 8 $\mu\text{g ml}^{-1}$ and mouse IL-7 (20 ng ml^{-1}) (Pharmingen), centrifuged at 1,800g for 45 min at room temperature, and incubated at 37 °C for 48 h. The GFP⁺ cells purified by fluorescence-activated cell sorting were transferred into deoxyguanosine (1.35 mM; Sigma)-treated (5 days) day 14 fetal thymi from MHC^{-/-} mice in the wells of a Terasaki plate (Applied Scientific). After 1 day, the repopulated thymus lobes were then cultured at 37 °C on the upper membrane surface of a Transwell culture well containing RPMI-1640 supplemented with 10% FCS. After 7 days culture, total cells were stained with Cychrome-conjugated anti-CD4 and PE-conjugated anti-CD8 monoclonal antibodies, then CD4⁺CD8⁺ cells were purified by electronic sorting. These CD4⁺CD8⁺ cells were reaggregated with TSC from C57BL/6 mice and cultured for 72 h. In the case of thymocytes from AND mice, total cells recovered after 7 days culture in repopulated thymic lobes were stimulated with Mu-DC and CD4⁺CD8⁺CD69⁺ cells were purified by electronic sorting, which were reaggregated with TSC from C57BL/6 mice and cultured for 60 h.

Received 23 September 1999; accepted 31 January 2000.

- Robey, E. & Fowlkes, B. J. Selective events in T cell development. *Annu. Rev. Immunol.* **12**, 675–705 (1994).
- Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
- von Boehmer, H. CD4/CD8 lineage commitment: back to instruction? *J. Exp. Med.* **183**, 713–715 (1996).
- Chan, S., Correia-Neves, M., Benoist, C. & Mathis, D. CD4/CD8 lineage commitment: matching fate with competence. *Immunol. Rev.* **165**, 195–207 (1998).
- Robey, E. *et al.* An activated form of Notch influences the choice between CD4 and CD8 T cell lineages. *Cell* **87**, 483–492 (1996).
- Deftos, M. L., He, Y. W., Ojala, E. W. & Bevan, M. J. Correlating notch signaling with thymocyte maturation. *Immunity* **9**, 777–786 (1998).
- Kaye, J. *et al.* Selective development of CD4⁺ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature* **341**, 746–749 (1989).
- Konig, R., Huang, L. Y. & Germain, R. N. MHC class II interaction with CD4 mediated by a region analogous to the MHC class I binding site for CD8. *Nature* **356**, 796–798 (1992).
- Riberdy, J. M., Mostaghel, E. & Doyle, C. Disruption of the CD4-major histocompatibility complex class II interaction blocks the development of CD4(+) T cells *in vivo*. *Proc. Natl Acad. Sci. USA* **95**, 4493–4498 (1998).
- Anderson, G., Jenkinson, E. J., Moore, N. C. & Owen, J. J. MHC class II-positive epithelium and mesenchyme cells are both required for T-cell development in the thymus. *Nature* **362**, 70–73 (1993).
- Lucas, B. & Germain, R. N. Unexpectedly complex regulation of CD4/CD8 coreceptor expression supports a revised model for CD4+CD8+ thymocyte differentiation. *Immunity* **5**, 461–477 (1996).
- Davis, C. B., Killeen, N., Crooks, M. E., Raulet, D. & Littman, D. R. Evidence for a stochastic mechanism in the differentiation of mature subsets of T lymphocytes. *Cell* **73**, 237–247 (1993).
- Veillette, A., Zuniga-Pflucker, J. C., Bolen, J. B. & Krusebeck, A. M. Engagement of CD4 and CD8 expressed on immature thymocytes induces activation of intracellular tyrosine phosphorylation pathways. *J. Exp. Med.* **170**, 1671–1680 (1989).
- Wiest, D. L. *et al.* Regulation of T cell receptor expression in immature CD4+CD8+ thymocytes by p56lck tyrosine kinase: basis for differential signaling by CD4 and CD8 in immature thymocytes expressing both coreceptor molecules. *J. Exp. Med.* **178**, 1701–1712 (1993).
- Kisielow, P., Bluthmann, H., Staerz, U. D., Steinmetz, M. & von Boehmer, H. Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature CD4+8+ thymocytes. *Nature* **333**, 742–746 (1988).
- Hogquist, K. A., Jameson, S. C. & Bevan, M. J. Strong agonist ligands for the T cell receptor do not mediate positive selection of functional CD8+ T cells. *Immunity* **3**, 79–86 (1995).
- Teh, H. S. *et al.* Thymic major histocompatibility complex antigens and the alpha beta T-cell receptor determine the CD4/CD8 phenotype of T cells. *Nature* **335**, 229–233 (1988).

- Basson, M. A., Bommhardt, U., Cole, M. S., Tso, J. Y. & Zamoyska, R. CD3 ligation on immature thymocytes generates antagonist-like signals appropriate for CD8 lineage commitment, independently of T cell receptor specificity. *J. Exp. Med.* **187**, 1249–1260 (1998).
- Merkenschlager, M. *et al.* How many thymocytes audition for selection? *J. Exp. Med.* **186**, 1149–1158 (1997).
- Iwata, M., Kuwata, T., Mukai, M., Tozawa, Y. & Yokoyama, M. Differential induction of helper and killer T cells from isolated CD4+CD8+ thymocytes in suspension culture. *Eur. J. Immunol.* **26**, 2081–2086 (1996).
- Grakoui, A. *et al.* The immunological synapse: a molecular machine controlling T cell activation. *Science* **285**, 221–227 (1999).
- Radtke, F. *et al.* Deficient T cell fate specification in mice with an induced inactivation of Notch1. *Immunity* **10**, 547–558 (1999).
- Marshall, C. J. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* **80**, 179–185 (1995).
- Matechak, E. O., Killeen, N., Hedrick, S. M. & Fowlkes, B. J. MHC class II-specific T cells can develop in the CD8 lineage when CD4 is absent. *Immunity* **4**, 337–347 (1996).
- Sharp, L. L., Schwarz, D. A., Bott, C. M., Marshall, C. J. & Hedrick, S. M. The influence of the MAPK pathway on T cell lineage commitment. *Immunity* **7**, 609–618 (1997).
- Bommhardt, U., Basson, M. A., Krummrei, U. & Zamoyska, R. Activation of the extracellular signal-related kinase/mitogen-activated protein kinase pathway discriminates CD4 versus CD8 lineage commitment in the thymus. *J. Immunol.* **163**, 715–722 (1999).
- Felli, M. P. *et al.* Expression pattern of notch1, 2 and 3 and jagged1 and 2 in lymphoid and stromal thymus components: distinct ligand-receptor interactions in intrathymic T cell development. *Int. Immunol.* **11**, 1017–1025 (1999).
- Beatus, P., Lundkvist, J., Oberg, C. & Lendahl, U. The notch 3 intracellular domain represses notch 1-mediated activation through Hairy/Enhancer of split (HES) promoters. *Development* **126**, 3925–3935 (1999).
- Inaba, M. *et al.* Distinct mechanisms of neonatal tolerance induced by dendritic cells and thymic B cells. *J. Exp. Med.* **173**, 549–559 (1991).
- Ouyang, W. *et al.* Inhibition of Th1 development mediated by GATA-3 through an IL-4-independent mechanism. *Immunity* **9**, 745–755 (1998).

Acknowledgements

This work was partially supported by grants from the Illinois Department of Public Health and from NIH to L.M. and American Cancer Society and the Arthritis Foundation to C.D. K.Y. was supported in part by a fellowship from the Japan Society for the Promotion of Science. We thank B.J. Fowlkes and R. H. Schwartz for their helpful comments on this manuscript, J. Delon for insightful suggestions concerning data interpretation, and M. Verma for generating the antisense Notch-1 construct.

Correspondence and requests for materials should be addressed to R.N.G. (e-mail: Ronald_Germain@nih.gov).

DNA repair protein Ku80 suppresses chromosomal aberrations and malignant transformation

Michael J. Difilippantonio*, Jie Zhu†, Hua Tang Chen†, Eric Meffre‡, Michel C. Nussenzweig‡, Edward E. Max§, Thomas Ried* & André Nussenzweig†

* Genetics Department, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

† Experimental Immunology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

‡ Laboratory of Molecular Immunology, and Howard Hughes Medical Institute, The Rockefeller Institute, New York, New York 10021, USA

§ Laboratory of Cell Regulation, Center for Biologics Evaluation and Research, Food and Drug Administration, National Institutes of Health, Bethesda, Maryland 20892, USA

Cancer susceptibility genes have been classified into two groups: gatekeepers and caretakers¹. Gatekeepers are genes that control cell proliferation and death, whereas caretakers are DNA repair genes whose inactivation leads to genetic instability. Abrogation of both caretaker and gatekeeper function markedly increases cancer susceptibility. Although the importance of Ku80 in DNA double-strand break repair is well established, neither Ku80 nor other components of the non-homologous end-joining pathway

are known to have a caretaker role in maintaining genomic stability. Here we show that mouse cells deficient for *Ku80* display a marked increase in chromosomal aberrations, including breakage, translocations and aneuploidy. Despite the observed chromosome instabilities, *Ku80*^{-/-} mice have only a slightly earlier onset of cancer^{2,3}. Loss of p53 synergizes with *Ku80* to promote tumorigenesis such that all *Ku80*^{-/-}*p53*^{-/-} mice succumb to disseminated pro-B-cell lymphoma before three months of age. Tumours result from a specific set of chromosomal translocations and gene amplifications involving IgH and c-Myc, reminiscent of Burkitt's lymphoma. We conclude that *Ku80* is a caretaker gene that maintains the integrity of the genome by a mechanism involving the suppression of chromosomal rearrangements.

The three subunits of the DNA-dependent protein kinase, Ku70, Ku80 and DNA-PKcs, are, together with XRCC4 and Ligase IV, essential for the non-homologous end-joining (NHEJ) pathway in mammalian cells. Mice that carry the targeted disruption of components of the NHEJ pathway share some common phenotypic features including arrested lymphocyte development and increased sensitivity to ionizing radiation³⁻⁹. Deficiencies in Ku80, Ku70, XRCC4 and Ligase IV, but not in DNA-PKcs, result in growth retardation and decreased proliferation *in vitro*. This senescence has been proposed to result from an inability to repair double-strand breaks (DSBs) in DNA incurred during normal DNA metabolism³.

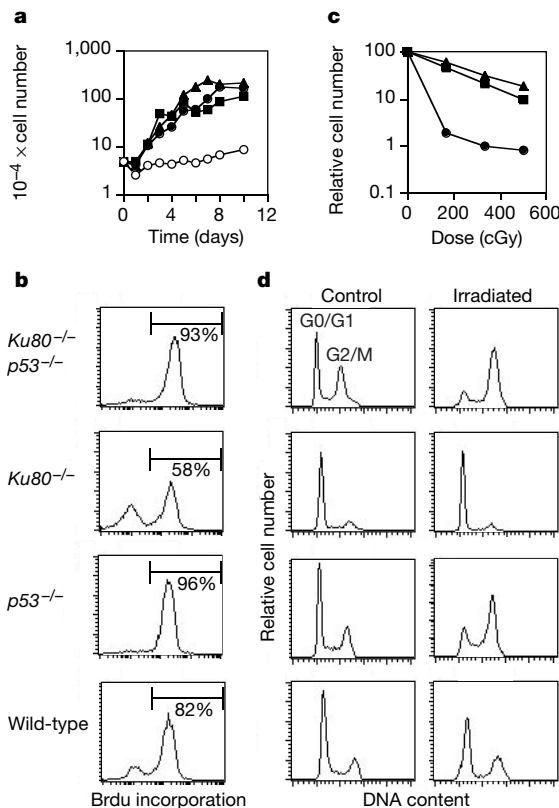


Figure 1 Growth characteristics of untreated and irradiated mouse embryo fibroblasts. **a**, Growth kinetics of *Ku80*^{-/-} (open circles), *p53*^{-/-} (filled triangles), *Ku80*^{-/-}*p53*^{-/-} (filled circles) and wild-type (filled squares) MEFs. **b**, Incorporation of BrdU in MEF cultures after 16 h of continuous labelling. **c**, Radiation sensitivity of *Ku80*^{-/-}*p53*^{-/-} (circles), *p53*^{-/-} (triangles) and wild-type (squares) MEFs, plotted as the fraction of surviving cells relative to unirradiated samples of the same genotype. *Ku80*^{-/-} MEFs exhibited a sevenfold decrease in cell count independent of the dose (83–600 cGy) owing to the premature senescence of unirradiated cultures, and the permanent arrest of the dividing population in response to γ -irradiation. **d**, DNA content histogram measured in untreated samples (control) and in cultures 24 h after treatment with 10 Gy of γ -radiation (irradiated).

p53 monitors chromosome damage and either arrests cell-cycle progression or triggers apoptosis in cells with unrepaired lesions¹⁰. To determine whether p53 is involved in the growth arrest of *Ku80*^{-/-} mice and mouse embryo fibroblasts (MEFs), we generated *Ku80*^{-/-}*p53*^{-/-} double-mutant mice. The absence of p53 did not markedly affect the initial viability of *Ku80*^{-/-}*p53*^{-/-} mice, and they were indistinguishable in size from *Ku80*^{-/-} littermates. By contrast, inactivation of p53 did alleviate the growth arrest of the *Ku80* knock-out MEFs (Fig. 1a, b). *Ku80*^{-/-}*p53*^{-/-} cells proliferated as rapidly and saturated at a similar density as *p53*^{-/-} cells (Fig. 1a), and contained a significantly greater cycling population than *Ku80*^{-/-} MEFs (Fig. 1b). Although the *Ku80*^{-/-} proliferation defect was abrogated by the concomitant loss of p53, *Ku80*^{-/-}*p53*^{-/-} MEFs were more sensitive to DNA damage induced by ionizing radiation than *p53*^{-/-} or wild-type controls (Fig. 1c). After γ -irradiation,

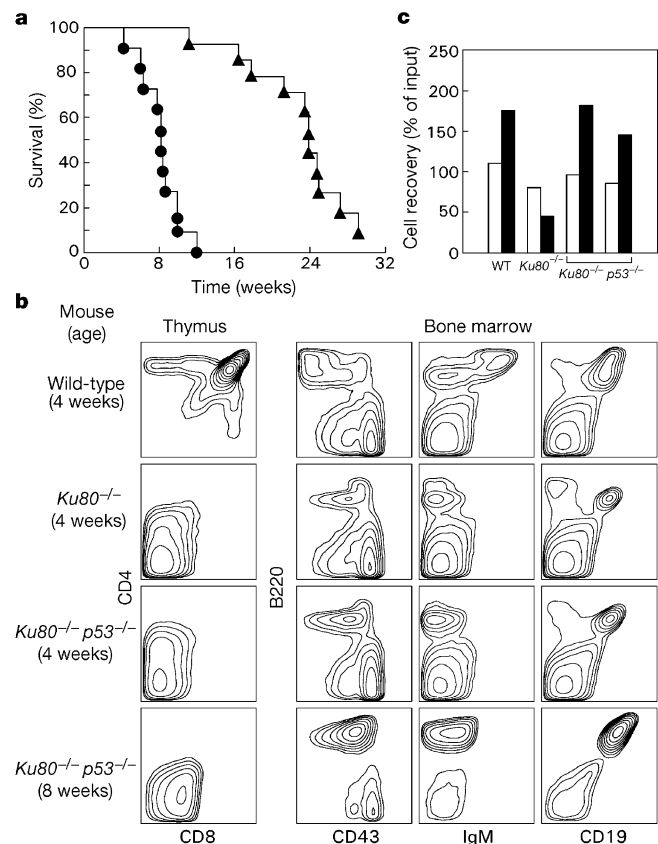


Figure 2 Analysis of lifespan and lymphocyte development. **a**, Kaplan–Meier analysis comparing mortality of *Ku80*^{+/+}*p53*^{-/-} (triangles) and *Ku80*^{-/-}*p53*^{-/-} (circles) mice as a function of time. The survival of *Ku80*^{-/-}*p53*^{-/-} mice is statistically different from that of the *Ku80*^{+/+}*p53*^{-/-} group; $P = 0.0001$. **b**, Flow cytometric analysis of thymocyte (left panel) and bone-marrow (right panels) suspensions obtained from wild-type, *Ku80*^{-/-} and *Ku80*^{-/-}*p53*^{-/-} mice. An 8-week-old *Ku80*^{-/-}*p53*^{-/-} mouse, which had a B-cell lymphoma, is compared with a 4-week-old non-tumour-bearing *Ku80*^{-/-}*p53*^{-/-} mouse, and with 4-week-old *Ku80*^{-/-} and wild-type littermates. Data are representative of 10 wild-type, 6 *Ku80*^{-/-}, 5 non-tumour-bearing *Ku80*^{-/-}*p53*^{-/-} mice 2–4 weeks old, and 12 *Ku80*^{-/-}*p53*^{-/-} animals that had tumours. Average thymus cellularity was 5×10^5 in *Ku80*^{-/-} mice, 10^6 in non-tumour-bearing *Ku80*^{-/-}*p53*^{-/-} mice and 2×10^8 in wild-type controls. Average number of B220⁺CD19⁺ cells in bone marrow was 4.6×10^5 in *Ku80*^{-/-} mice, 9.6×10^5 in non-tumour-bearing *Ku80*^{-/-}*p53*^{-/-} mice and 2.8×10^6 for wild-type controls. The immuno-phenotypes of *Ku80*^{+/+}*p53*^{-/-} and wild-type mice were similar (not shown). **c**, Survival of wild-type (WT) B220⁺CD43⁺IgM⁺ pro-B cells, and *Ku80*^{-/-}*p53*^{-/-} and *Ku80*^{-/-} CD19⁺ B cells after 1 d (open columns) and 3 d (filled columns) in culture, plotted relative to input cell number at day 0. Results from two independent *Ku80*^{-/-}*p53*^{-/-} mice are plotted.

Ku80^{-/-}*p53*^{-/-} MEFs initially accumulated in the G2/M phase of the cell cycle (Fig. 1d). Within 48–96 hours after irradiation, *Ku80*^{-/-}*p53*^{-/-} MEFs lost their adherence to the tissue culture dish, exhibited extensive nuclear fragmentation and showed an increase in the number of cells with a DNA content less than that in G0/G1 (not shown). By contrast, *Ku80*^{-/-} MEFs were arrested permanently in the G0/G1 and G2/M phases of the cell cycle in response to γ -irradiation³ (Fig. 1d). Thus, p53 is required for early senescence in *Ku80*^{-/-} fibroblasts, indicating that cell-cycle arrest in these cells is correlated with an accumulation of DNA damage (see below). The fact that a loss of p53 does not rescue the size of *Ku80*^{-/-} mice indicates either that the dwarfism is unrelated to DNA damage or that *Ku80*^{-/-}*p53*^{-/-} cells that sustain irreparable damage are eliminated by p53-independent mechanisms *in vivo*.

Although *Ku80*^{-/-}*p53*^{-/-} mice developed normally, all (*n* = 11) such mice died within 12 weeks after birth (Fig. 2a) from disseminated lymphoma. Tumour cells expressed B-cell lineage-specific surface markers B220, CD19 and CD43, but were negative for IgM, indicating maturational arrest at the pro-B-cell stage of development (Fig. 2b; right panel, *Ku80*^{-/-}*p53*^{-/-} 8 weeks). In contrast, *p53*^{-/-} mice are predisposed to thymic lymphomas¹¹, which develop at a slower rate than *Ku80*^{-/-}*p53*^{-/-} pro-B-cell lymphomas (Fig. 2a), and *Ku80*^{-/-} mice only occasionally develop lymphoma² (only 1 out of 11 *Ku80*^{-/-} mice developed T-cell lymphoma (Thy1.2⁺, CD4⁺, CD8⁺) after seven months). We conclude that the loss of p53 in the *Ku80* knockout background invariably results in the rapid onset of B-lineage tumours before the age at which T-lineage tumours occasionally arise in *Ku80*^{-/-} mice.

The shift from T-cell to B-cell lymphomas in *Ku80*^{-/-}*p53*^{-/-} mice was not accompanied by an obvious alteration in T- or B-lineage surface markers. *Ku80*^{-/-}*p53*^{-/-} thymocytes resembled *Ku80*^{-/-} thy-

mocytes in that they developed no further than the CD4⁺CD8⁻ stage (Fig. 2b; left panel), whereas B cells remained arrested at the B220⁺, CD19⁺, CD43⁺, IgM⁻ pro-B-cell stage (Fig. 2b; right panel). However, the loss of p53 had different effects on the survival of T-cell and B-cell precursors. Whereas both *Ku80*^{-/-} pro-B and pro-T cells showed high susceptibility to apoptosis (as indicated by annexin staining; data not shown) and poor proliferation, loss of p53 from the double knockouts seemed to rescue *Ku80*^{-/-} pro-B cells, but not pro-T cells (not shown), allowing normal proliferation *in vitro* (Fig. 2c). We conclude that p53 is required for the induction of apoptosis in B-lymphocyte but not T-lymphocyte precursors that harbour chromosome abnormalities induced by the loss of *Ku80*. This difference might in part explain the increased susceptibility of *Ku80*^{-/-}*p53*^{-/-} mice to B-lineage malignancies.

The observation that *Ku80*^{-/-}*p53*^{-/-} mice develop lymphomas at an accelerated rate suggests that Ku80 and p53 normally cooperate to limit the oncogenic potential of DSBs during assembly of the antigen receptor gene. Chromosome analysis by spectral karyotyping (SKY)¹² revealed karyotype heterogeneity within each *Ku80*^{-/-}*p53*^{-/-} tumour (*n* = 4), exemplified by translocations and apparently random gains or losses of whole chromosomes (Fig. 3a). Metaphases from 20 identically treated *Rag2*^{-/-} pro-B cells exhibited no detectable aberrations. All tumours contained a translocation of chromosome 12, on which the immunoglobulin heavy chain (IgH) is located (Fig. 3b; SKY). In three tumours (T2, T3 and T4) the partner was chromosome 15, whereas in the remaining tumour (T1) chromosome 3 was the fusion partner (Fig. 3b; SKY). The one *Ku80*^{-/-} T-cell lymphoma that arose during the 7-month observation period exhibited a translocation that fused chromosomes 3 and 7 (not shown). Thus, both *Ku80*^{-/-} and *Ku80*^{-/-}*p53*^{-/-} mice develop lymphomas that exhibit chromosome translocations.

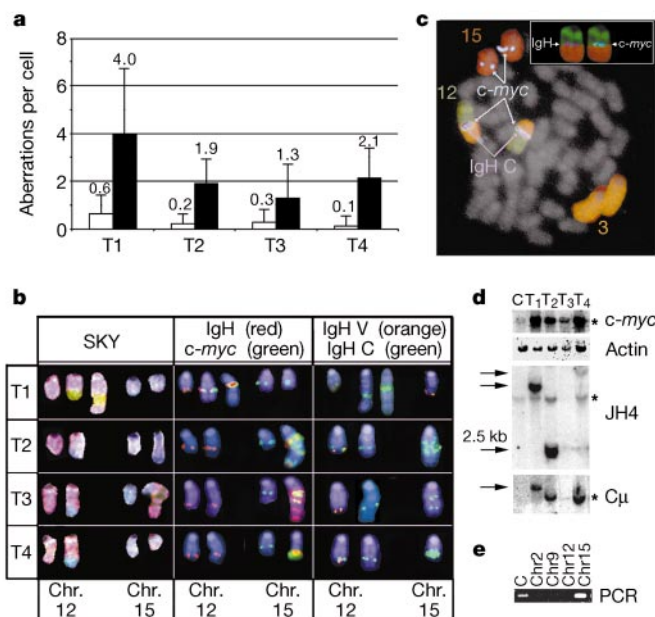


Figure 3 *c-myc* and IgH-associated rearrangements and amplifications in *Ku80*^{-/-}*p53*^{-/-} tumours. **a**, Average number of translocations (other than those juxtaposing chromosomes 12 and 15), unshaded, or gains and losses of chromosomes, shaded, in *Ku80*^{-/-}*p53*^{-/-} tumours (T1–T4). **b**, SKY analysis revealed translocations involving chromosomes (Chr.) 12 (pink) and 15 (blue) in *Ku80*^{-/-}*p53*^{-/-} tumours 2–4 (T2–T4) and 12 and 3 (green) in T1. FISH with probes for IgH C_{H1} (red) and c-myc (green) reveal co-localization and co-amplification of these genes as well as signals on the normal chromosomes. Analysis with probes for IgH variable (orange) and constant (green) regions show co-localization on the normal chromosome 12. The absence of the variable region

on the derivative 12 shows that this translocation disrupts the IgH locus. Chromosomes in the middle and right panels were identified by hybridization with painting probes (as in **c**) and are counterstained with 4,6-diamidino-2-phenylindole (DAPI; blue). **c**, Five-colour FISH analysis of a metaphase from tumour T1. The inset shows the localization of individual c-myc and IgH signals at the breakpoint between chromosomes 12 and 3. **d**, Southern blot analysis showing germline bands (stars) as well as rearrangements (arrows) and/or amplifications of c-myc and IgH. **e**, PCR amplification of tumour 2 breakpoint indicates that the JH1 fusion partner originated from chromosome 15. Abbreviations: C: Liver; T1–T4: *Ku80*^{-/-}*p53*^{-/-} tumors 1–4.

Translocations between chromosomes 12 and 15 in *Ku80*^{-/-}*p53*^{-/-} lymphomas resemble the translocations in mouse plasmacytomas and Burkitt's lymphomas that juxtapose the *c-myc* oncogene to the IgH locus¹³. Translocations involving the IgH locus, but not *c-myc*, have been reported in B-cell lymphomas arising from crosses between *SCID* (severe combined immunodeficiency) mice—which are defective in DNA-PKcs—and *p53*^{-/-} mice¹⁴. We therefore performed fluorescence *in situ* hybridization (FISH) analysis to determine whether *c-myc* and IgH genes were translocated in *Ku80*^{-/-}*p53*^{-/-} lymphomas. All *Ku80*^{-/-}*p53*^{-/-} lymphomas showed colocalization of *c-myc* and IgH with an apparent amplification of both genes (Fig. 3b; IgH, *c-myc*). In the one *Ku80*^{-/-}*p53*^{-/-} tumour that showed a 12:3 translocation, *c-myc* and IgH co-localized precisely at the breakpoint between chromosomes 12 and 3, indicating a three-way (12:15:3) translocation (Fig. 3c). Amplification of *c-myc* and IgH was confirmed by Southern blot analysis (Fig. 3d), and IgH locus rearrangements were found in all tumours (Fig. 3d). Furthermore, FISH analysis mapped the breakpoint on chromosome 12 to between the IgH variable cluster and constant region (Fig. 3b; IgH V, IgH C). We conclude that *Ku80*^{-/-}*p53*^{-/-} lymphomas result from a specific set of chromosomal translocations and amplifications involving *c-myc* and IgH, and therefore differ from *SCIDxp53*^{-/-} B-cell lymphomas¹⁴. DNA-PKcs-independent functions of Ku, as established in previous studies, might contribute to the cytogenetic differences in these tumours^{3,15}.

To determine the precise nature of the translocations in *Ku80*^{-/-}*p53*^{-/-} lymphomas, we cloned the novel JH-hybridizing *Eco*RI fragment from one of the four tumours (Fig. 3d; T2). Sequence analysis revealed that the 2.5-kilobase (kb) fragment was a fusion of JH1 to a 173-base pair (bp) segment that bore no significant homologies to previously reported sequences. To determine the chromosomal origin of the 173-bp DNA fragment, FACS-sorted chromosomes were screened by PCR (Fig. 3e). Unfractionated liver DNA and purified chromosome 15 DNA showed a signal, whereas chromosomes 2, 9 and 12 were negative (Fig. 3e). Loss of the variable region on the der(12) chromosome and the proximity of the translocation to JH1 suggest that the translocation resulted from aberrant repair of DNA breaks produced in pro-B cells during V(D)J recombination.

One of the hallmarks of malignant transformation is genomic instability, which promotes a wide range of mutations, including chromosome deletions, gene amplifications, translocations and

polyploidy¹⁶. To determine whether loss of *Ku80* results in chromosome instability, we karyotyped metaphases from passage-1 MEFs (Fig. 4). Whereas only 9% of metaphases from wild-type mice were karyotypically abnormal, we observed profound aberrations in the integrity of the *Ku80*^{-/-} genome (Fig. 4a); 83% of *Ku80*^{-/-} metaphases had breaks or translocations (Fig. 4a) and 15% displayed polyploidy (that is, 3*n*, 4*n* or 5*n* chromosome content, where 2*n* reflects a normal diploid complement). Although chromosome aberrations have been reported in primary dermal fibroblasts from *Ku80*^{-/-}, *Ku80*^{+/-} and ligase IV^{+/-} mice¹⁷, we found no significant differences between *Ku80*^{+/-} and wild-type MEFs. The pattern of damage in the *Ku80*^{-/-} metaphases was further characterized by SKY, which showed chromatid breaks, acentric fragments, triradials, dicentrics, duplications and asymmetric exchanges (Fig. 4b). *Ku80*^{-/-}*p53*^{-/-} MEFs were similar to *Ku80*^{-/-}, whereas *p53*^{-/-} MEFs had slightly fewer aberrations (Fig. 4a). Thus, the absence of Ku80 precipitates genomic instability, which is not augmented by the additional loss of *p53*.

An apparently random gain or loss of chromosomes also occurred in 70% of *Ku80*^{-/-} MEFs (not shown). Because aneuploidy has been associated with centrosome amplification in *BRCA1*^{-/-} and *p53*^{-/-} MEFs^{18,19}, we looked at these structures in early-passage MEFs. Our analysis of γ -tubulin structures revealed a similar centrosome amplification in both *Ku80*^{-/-}*p53*^{-/-} and *p53*^{-/-} MEFs. However, despite the numerical aberrations seen in *Ku80*^{-/-} MEFs, centrosomes in these cells were indistinguishable from normal fibroblast controls (Fig. 4c; not shown). Thus, the acquisition of numerical chromosome aberrations in *Ku80*^{-/-} MEFs occurs by a mechanism unrelated to centrosome amplification.

Unresolved signal ends in *Ku80*^{-/-} lymphocytes might contribute to the development of translocations by acting as transposable elements^{20,21}. Alternatively, gross chromosomal rearrangements in *Ku80*^{-/-} mice might result from the absence of Ku-mediated tethering of broken ends, which normally facilitates their subsequent ligation^{22,23}. Promiscuous chromosome fusions in the absence of Ku80 could either be accomplished by other components of the apparatus of the NHEJ pathway (that is, Ku70, DNA-PKcs, XRCC4 or Ligase IV) or involve alternative DNA repair pathways such as single-strand annealing or break-induced replication²⁴. In either case, the finding that Ku80 suppresses gross genomic rearrangements places Ku80 in the class of caretaker genes, which include *ATM*, *BRCA1* and several proteins that function in homologous

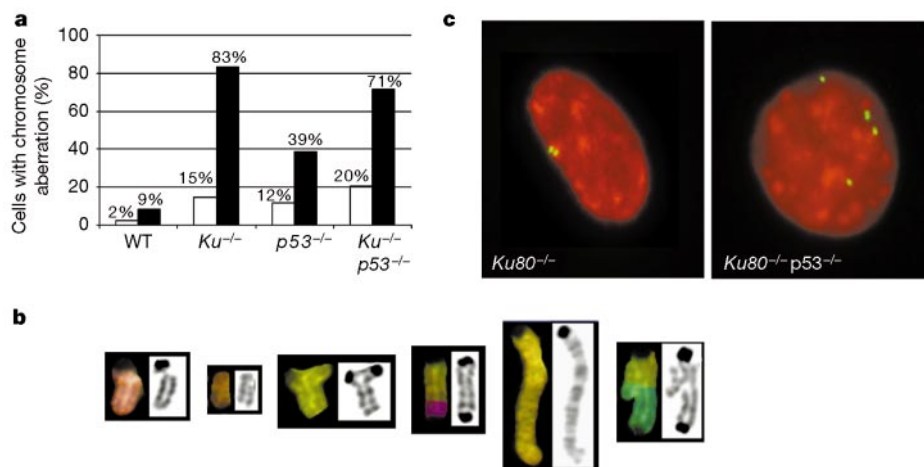


Figure 4 Chromosome aberrations in *Ku80*^{-/-} and *Ku80*^{-/-}*p53*^{-/-} MEFs. **a**, Percentage of cells with DNA breaks or translocations, shaded (*n* = 35), and polyploidy, unshaded (*n* = 400), observed in passage 1 MEFs. WT, wild type. **b**, Typical chromosome aberrations found in *Ku80*^{-/-} MEFs. From left to right, chromatid break, acentric fragment, tri-radial,

dicentric, duplication and an asymmetric exchange. **c**, γ -Tubulin staining for centrosomes (green) in *Ku80*^{-/-} (left) and *Ku80*^{-/-}*p53*^{-/-} (right) MEFs counterstained with propidium iodide (red).

recombination^{1,25,26}. p53 provides an additional safeguard against the genomic instability inherent in Ku80-deficient cells by either arresting cell division (as in *Ku80*^{-/-} MEFs) or triggering apoptosis (as in *Ku80*^{-/-} B lymphocytes). Consistent with this model is the observation that the abrogation of both Ku80 caretaker and p53 gatekeeper functions markedly increases malignant transformation, especially in lymphocytes that undergo physiological DSBs during the course of V(D)J recombination. Whereas the phenotype of *Ku80*^{-/-} mice is more severe than that of *DNA-PKcs*^{-/-} mice, we would expect that other components of the NHEJ machinery (Ku70, XRCC4 and Ligase IV) also have an essential role in maintaining genomic stability. Thus, mutation of the NHEJ pathway components might facilitate the development of malignancy by increasing the frequency of chromosomal aberrations. □

Methods

Generation and screening of mice

Ku80 mutant mice were generated from *Ku80*^{+/-} intercrosses and screened by polymerase chain reaction (PCR) as described³. *p53*^{-/-} mice were obtained from Taconic Laboratories, and *p53* genotyping was performed as described^{27,28}. Mice heterozygous for *Ku80* and *p53* were crossed to obtain progeny homozygous for both *Ku80*-null and *p53*-null alleles. To generate mouse embryo fibroblasts, *Ku80*^{+/-}*p53*^{+/-} males were crossed with *Ku80*^{+/-}*p53*^{+/-} females and embryos were isolated 13.5 days after plug formation³.

Growth and irradiation assays

Passage-1 MEFs (10⁵) were plated in duplicate six-well dishes and the medium was replaced daily. Individual dishes were treated with trypsin, counted and averaged. For continuous labelling experiments, 10⁵ cells were plated with 50 μ M bromodeoxyuridine (BrdU) and grown for 16 h. Cells were stained with fluorescein isothiocyanate (FITC)-conjugated anti-BrdU antibodies and propidium iodide (PI) in accordance with the manufacturer's specifications (Becton Dickinson). For radiation survival experiments, cells were exposed to graded doses of γ -rays, counted 4 d after irradiation, and normalized to the number of cells in unirradiated controls of the same genotype. Aliquots fixed at 24–96 h after irradiation were stained with PI for measurements of DNA content.

Spectral karyotyping, FISH and immunocytochemistry

MEFs from wild-type, *p53*^{-/-}, *Ku80*^{-/-} and *Ku80*^{-/-}*p53*^{-/-} mice were cultured in DMEM containing 10% fetal bovine serum (FBS). Suspensions of tumour cells derived from lymph node were grown in RPMI containing 10% FBS, 50 units ml⁻¹ interleukin-7 (IL-7) (R&D Systems) and 20 units ml⁻¹ IL-2 (Boehringer Mannheim). Tumour T1 was analysed after one passage in culture, whereas T2, T3 and T4 were primary tumours that were processed immediately for SKY, FISH and Southern blot analysis. MEFs and tumour cells were arrested at mitosis by treatment with Colcemid (Gibco/BRL) at 0.1 μ g ml⁻¹ for 0.5–3 h. Mitotic chromosome spreads were prepared by following standard procedures, and SKY analysis was performed as described¹². For FISH analysis, metaphases were hybridized with BAC probes for the following regions: IgH constant region from C₁ to 3' of C₂, IgH variable region representing the distal end of the V cluster (VJ588), and *c-myc*. Flow-sorted single chromosomes were used for painting probes. Immunostaining of centrosomes with γ -tubulin antibodies was performed as described²⁹.

Flow cytometry

Single-cell suspensions from thymocytes, bone marrow, lymph node and tumours were stained for various lineage-specific surface markers as described³. The following monoclonal antibodies (PharMingen) were used in various two-colour combinations: phycoerythrin (PE)-conjugated anti-CD4 (clone RM4-5), FITC-conjugated anti-CD8a (clone 53-6.7), FITC-conjugated anti-T-cell antigen receptor (clone H57-597), PE-conjugated anti-B220 (clone RA3-6B2), FITC-conjugated anti-CD19 (clone 1D3), FITC-conjugated anti-IgM (clone II/41), FITC-conjugated anti-CD43 (clone S7) and PE-conjugated Thy1.2 (clone 53-2.1). At least 100,000 data points were collected on a modified immunocytometry systems FACSSTAR Plus using CellQuest software (Becton Dickinson).

Lymphocyte purification and culture

Bone marrow B cells from mutant and wild-type mice were enriched by positive selection with MACS mouse CD19 microbeads (Miltenyi Biotec). Wild-type B cells were further sorted into B220⁺CD43⁺CD25⁺IgM⁺ pro-B cells. Pro-B cells were co-cultured for 3 d with irradiated S17 stromal cells and mouse rIL-7 (10 ng ml⁻¹). Wild-type CD4⁺CD8⁺ T-cell precursors were enriched by negative selection with magnetic beads coated with anti-mouse CD4 and anti-mouse CD8 (Dyna). Apoptosis and viability of B and T cells were analysed with PE-Annexin (PharMingen) and staining with PI.

Southern blot analysis and digestion–circularization PCR

Genomic DNA (10 μ g) from *Ku80*^{-/-}*p53*^{-/-} tumours was digested with *Eco*RI. Southern blots were sequentially hybridized to ³²P-labelled probes for *c-myc* and mouse actin cDNA;

a duplicate blot was prepared and sequentially hybridized with a 1.1-kb JH4 and a 900-bp C μ genomic fragment. For digestion–circularization PCR³⁰, 2 μ g of genomic DNA was digested with *Eco*RI and ligated under dilute conditions that promote self-ligation³⁰. Primers JH4F (5'-CTTCCCCAAATAGCTTGCC-3') and JH4R (5'-CCCACCAAAACCGAAAGTCCA-3') were used to amplify the 2.5-kb product by using the Expand High Fidelity PCR System (Boehringer). The amplified product was cloned into PCR-XL-TOPO (Invitrogen) and individual clones were isolated for sequencing. PCR primers PTK2F (5'-TGAGTTTCTGGGTGTGGGC-3') and PTK2R (5'-GCATAGGATTCGGCGTAGA-3'), derived from the 173-bp sequence fused to JH1, were used to screen 10 ng of DNA derived from liver and flow-sorted mouse chromosomes 2, 9, 12 and 15.

Received 4 January; accepted 22 February 2000.

- Kinzler, K. W. & Vogelstein, B. Gatekeepers and caretakers. *Nature* **386**, 761–763 (1997).
- Vogel, H., Lim, D. S., Karsenty, G., Finegold, M. & Hasty, P. Deletion of Ku86 causes early onset of senescence in mice. *Proc. Natl Acad. Sci. USA* **96**, 10770–10775 (1999).
- Nussenzweig, A. *et al.* Requirement for Ku80 in growth and immunoglobulin V(D)J recombination. *Nature* **382**, 551–555 (1996).
- Zhu, C., Bogue, M. A., Lim, D. S., Hasty, P. & Roth, D. B. Ku86-deficient mice exhibit severe combined immunodeficiency and defective processing of V(D)J recombination intermediates. *Cell* **86**, 379–389 (1996).
- Ouyang, H. *et al.* Ku70 is required for DNA repair but not for T cell antigen receptor gene recombination *in vivo*. *J. Exp. Med.* **186**, 921–929 (1997).
- Gu, Y. *et al.* Growth retardation and leaky SCID phenotype of Ku70-deficient mice. *Immunity* **7**, 367–376 (1997).
- Gao, Y. *et al.* A critical role for DNA end-joining proteins in both lymphogenesis and neurogenesis. *Cell* **95**, 891–902 (1998).
- Frank, K. M. *et al.* Late embryonic lethality and impaired V(D)J recombination in mice lacking DNA ligase IV. *Nature* **396**, 173–177 (1998).
- Barnes, D. E., Stamp, G., Rosewell, I., Denzel, A. & Lindahl, T. Targeted disruption of the gene encoding DNA ligase IV leads to lethality in embryonic mice. *Curr. Biol.* **8**, 1395–1398 (1998).
- Levine, A. J. p53, the cellular gatekeeper for growth and division. *Cell* **88**, 323–331 (1997).
- Jacks, T. *et al.* Tumor spectrum analysis in p53-mutant mice. *Curr. Biol.* **4**, 1–7 (1994).
- Liyanage, M. *et al.* Multicolour spectral karyotyping of mouse chromosomes. *Nature Genet.* **14**, 312–315 (1996).
- Taub, R. *et al.* Translocation of the *c-myc* gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. *Proc. Natl Acad. Sci. USA* **79**, 7837–7841 (1982).
- Vanase, G. J. *et al.* Genetic pathway to recurrent chromosome translocations in murine lymphoma involves V(D)J recombination. *J. Clin. Invest.* **103**, 1669–1675 (1999).
- Gao, Y. *et al.* A targeted DNA-PKcs-null mutation reveals DNA-PK-independent functions for KU in V(D)J recombination. *Immunity* **9**, 367–376 (1998).
- Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
- Karanjawa, Z. E., Grawunder, U., Hsieh, C. L. & Lieber, M. R. The nonhomologous DNA end joining pathway is important for chromosome stability in primary fibroblasts. *Curr. Biol.* **9**, 1501–1504 (1999).
- Fukasawa, K., Choi, T., Kuriyama, R., Rulong, S. & Vande Woude, G. F. Abnormal centrosome amplification in the absence of p53. *Science* **271**, 1744–1747 (1996).
- Xu, X. *et al.* Centrosome amplification and a defective G2-M cell cycle checkpoint induce genetic instability in BRCA1 exon 11 isoform-deficient cells. *Mol. Cell* **3**, 389–395 (1999).
- Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by Rag1 and Rag2 and its implications for the evolution of the immune system. *Nature* **394**, 744–751 (1998).
- Hiom, K., Melek, M. & Gellert, M. DNA transposition by the Rag1 and Rag2 proteins: a possible source of oncogenic translocations. *Cell* **94**, 463–470 (1998).
- Ramsden, D. A. & Gellert, M. Ku protein stimulates DNA end joining by mammalian DNA ligases: a direct role for Ku in repair of DNA double-strand breaks. *EMBO J* **17**, 609–614 (1998).
- Smith, G. C. M. & Jackson, S. P. The DNA-dependent protein kinase. *Genes Dev.* **13**, 916–934 (1999).
- Chen, C., Umez, K. & Kolodner, R. D. Chromosomal rearrangements occur in *S. cerevisiae* rfa1 mutator mutants due to mutagenic lesions processed by double-strand-break repair. *Mol. Cell* **2**, 9–22 (1998).
- Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. Brca1 controls homology-directed DNA repair. *Mol. Cell* **4**, 511–518 (1999).
- Johnson, R. D., Liu, N. & Jasin, M. Mammalian XRCC2 promotes the repair of DNA double-strand breaks by homologous recombination. *Nature* **401**, 397–399 (1999).
- Timme, T. L. & Thompson, T. C. Rapid allele analysis of p53 knockout mice. *Biotechniques* **17**, 462–463 (1994).
- Donehower, L. A. *et al.* Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* **356**, 215–221 (1992).
- Ghadimi, B. M. *et al.* Centrosome amplification and instability occurs exclusively in aneuploid, but not in diploid colorectal cancer cell lines, and correlates with numerical chromosomal aberrations. *Genes Chromosomes Cancer* **27**, 183–190 (2000).
- Chu, C. C., Paul, W. E. & Max, E. E. Quantitation of immunoglobulin μ - γ 1 heavy chain switch region recombination by a digestion–circularization polymerase chain reaction method. *Proc. Natl Acad. Sci. USA* **89**, 6978–6982 (1992).

Acknowledgements

We thank A. Singer, R. Hodes, A. Bhandoola, E. Besmer and S. Sharrow for comments on the manuscript and helpful discussions; K. Huppi, B. Malynn and R. Riblet for probes; and D. Liewehr, S. Steinberg and T. Brotz for assistance. M.C.N. is an investigator of the Howard Hughes Medical Institute. A.N. was supported in part by an award from the Arthritis Foundation.

Correspondence and requests for materials should be addressed to A.N. (e-mail: andre_nussenzweig@nih.gov).

Yeast Sm-like proteins function in mRNA decapping and decay

Sundaresan Tharun^{*,†}, Weihai He^{*,†}, Andrew E. Mayes^{‡,§}, Pascal Lennertz^{*}, Jean D. Beggs[‡] & Roy Parker^{*}

^{*}Department of Molecular and Cellular Biology and Howard Hughes Medical Institute, University of Arizona, Tucson, Arizona 85721, USA

[‡]Institute of Cell and Molecular Biology, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, UK

[§]Unilever Research Colworth, Colworth House, Sharnbrook, Bedford, MK44 1LQ, UK

[†]These authors contributed equally to this work

One of the main mechanisms of messenger RNA degradation in eukaryotes occurs by deadenylation-dependent decapping which leads to 5'-to-3' decay^{1,2}. A family of Sm-like (Lsm) proteins has been identified, members of which contain the 'Sm' sequence motif, form a complex with U6 small nuclear RNA and are required for pre-mRNA splicing³⁻⁹. Here we show that mutations in seven yeast Lsm proteins (Lsm1–Lsm7) also lead to inhibition of mRNA decapping. In addition, the Lsm1–Lsm7 proteins co-immunoprecipitate with the mRNA decapping enzyme (Dcp1), a decapping activator (Pat1/Mrt1) and with mRNA. This indicates that the Lsm proteins may promote decapping by interactions with the mRNA and the decapping machinery. In addition, the Lsm complex that functions in mRNA decay appears to be distinct from the U6-associated Lsm complex, indicating that Lsm proteins form specific complexes that affect different aspects of mRNA metabolism.

We proposed that the family of Lsm proteins affects mRNA degradation because *lsm1* strains are deficient in mRNA decapping¹⁰ and the Lsm proteins interact with each other^{6,7}. Thus, we examined the decay of the *MFA2pG* and *PGK1pG* reporter mRNAs¹¹ in strains with the nonessential *LSM1*, *LSM6* or *LSM7* genes (each of which is thermosensitive) deleted, and in strains carrying temperature-sensitive forms of *LSM2* or *LSM5*, or a partial loss-of-function allele of *LSM8*. The reporter mRNAs are under the control of the *GAL1* UAS to allow measurement of their decay rates and contain a poly(G) tract in their 3' untranslated region (UTR), which traps decay intermediates by blocking the 5'-to-3' exonuclease in *cis*^{12,13}. Mutations inhibiting mRNA turnover accumulate full-length mRNA and show a decrease in the levels of the decay intermediate¹².

Two observations indicated that loss-of-function alleles of *LSM1–LSM7* inhibited mRNA degradation. First, loss-of-function alleles in the *LSM1*, *LSM2*, *LSM5*, *LSM6* and *LSM7* genes led to increased levels of full-length mRNA and decreased levels of decay intermediate for both the *MFA2pG* (Fig. 1, in 0 min) and *PGK1pG* (data not shown) transcripts. Second, the *MFA2pG* transcript showed a slower decay rate (Fig. 1) in the *lsm1*, *lsm2*, *lsm5*, *lsm6* and *lsm7* mutant strains as compared with wild-type strains. Although the *lsm1*, *lsm2*, *lsm5*, *lsm6* and *lsm7* alleles cause thermosensitive growth, we observed the effects on mRNA decay at both 24°C (Fig. 1) and 37°C (data not shown). The *lsm8-1* allele examined did not affect mRNA turnover. The inactivation of the Lsm proteins caused a partial defect in mRNA decay as compared with the complete block seen in *dcp1Δ* and *dcp2Δ* strains^{14,15}, suggesting that Lsm proteins activate mRNA decay. Consistent with this view, an *lsm1Δ lsm6Δ* double mutant strain did not show an increased defect in mRNA decay (Fig. 1).

To determine which step in mRNA degradation was affected in the *lsm* mutants, we examined the structure of the *MFA2pG* mRNA in these strains. Strains deficient in decapping or in 5'-to-3' exonucleolysis accumulate oligoadenylated, full-length mRNA species, with or without cap, respectively^{10,12,14–16}. Thus, we

examined the poly(A) distribution on the *MFA2pG* mRNA in the *lsm* mutants with or without fractionation into capped and uncapped fractions by immunoprecipitation with anti-cap antibodies. Strains mutant in *LSM1–LSM7* accumulated *MFA2pG* transcripts with short poly(A) tails (Fig. 2, T lanes), indicating that these strains were deficient in decapping or in 5'-to-3' exonucleolysis. A substantial amount of the accumulated oligoadenylated species in the *lsm1*, *lsm2*, *lsm5*, *lsm6* and *lsm7* strains was immunoprecipitated by antisera against the cap structure, similar to the *dcp1Δ* strain, wherein decapping is blocked (Fig. 2). This indicated that *LSM2*, *LSM5*, *LSM6* and *LSM7* are required, as is *LSM1* (ref. 10) for efficient mRNA decapping. Moreover, in strains where the only copy of the essential *LSM3* or *LSM4* is expressed from a *GAL* promoter, transfer of the culture to glucose medium (thereby depleting Lsm3 or Lsm4 (ref. 7)) led to the accumulation of capped, deadenylated *MFA2pG* mRNAs (Fig. 2). This observation, together with the co-immunoprecipitation of Lsm3 and Lsm4 with mRNA decay factors (see below), suggests that *LSM3* and *LSM4* are also required for efficient mRNA decapping.

Two-hybrid analysis has identified interactions between Lsm proteins and the Dcp1 decapping enzyme, the Xrn1 5'-to-3' exonuclease and Dcp2 (ref. 17). The Lsm proteins also show

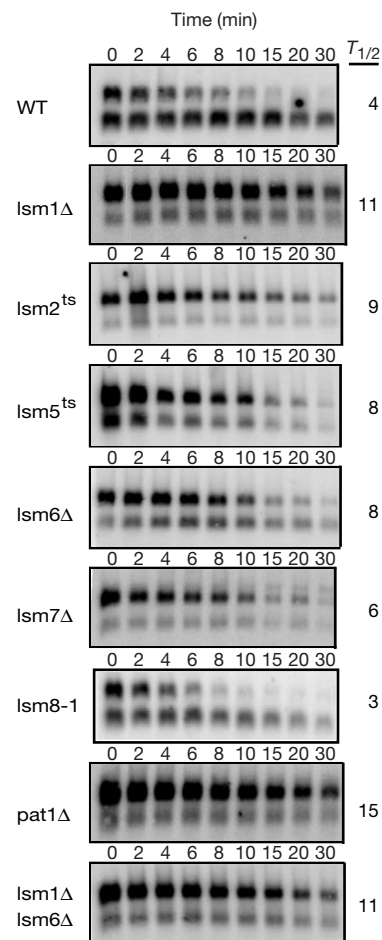


Figure 1 *MFA2pG* mRNA is stabilized in most *lsm* mutant strains. The decay of the *MFA2pG* transcript after repression of transcription by the addition of glucose in *lsm* mutants and wild-type (WT) strain at 24°C is shown. The top band is the full-length mRNA and the lower band is the mRNA fragment produced by 5'-to-3' exonucleolytic digestion to the 5' side of the poly(G) tract¹¹. The strains used were WT (yRP841), *lsm1Δ* (yRP1365), *lsm2ts* (yRP1366), *lsm5ts* (yRP1367), *lsm6Δ* (yRP1369), *lsm7Δ* (yRP1370), *lsm8-1*(BP4), *pat1/mrt1Δ* (yRP1372) and *lsm1Δ lsm6Δ* (yRP1407).

two-hybrid interactions with Pat1 (YCR077c)¹⁷, a protein originally identified through a two-hybrid interaction with topoisomerase II (ref. 18). *Pat1Δ* strains have multiple phenotypes, including temperature sensitivity, that are similar to mutations in the *MRT1* gene, an uncloned locus known to affect mRNA decapping¹². There is evidence that *MRT1* and *PAT1* are the same gene. First, the *PAT1* gene complements the temperature sensitivity and mRNA decay defect in *mrt1* strains (data not shown). Second, *pat1Δ* strains show a defect in mRNA decay similar to the *mrt1* alleles (Fig. 1). Third, sequencing of the *PAT1* gene from strains carrying the *mrt1-1*, *mrt1-2* and *mrt1-3* alleles identified mutations creating nonsense codons at amino acids 547, 157 and 308, respectively. Fourth, the *mrt1* alleles and the *pat1Δ* are genetically linked and *pat1Δ* fails to complement the *mrt1* lesions in diploids (data not shown). Thus, *PAT1/MRT1* is required for efficient mRNA decapping.

To investigate the significance of the two-hybrid interactions between the Lsm proteins and the proteins required for mRNA decapping, we immunoprecipitated epitope-tagged proteins Lsm1–

Lsm8 and determined whether Dcp1 or Pat1/Mrt1 co-immunoprecipitated by using western analysis. Dcp1 co-immunoprecipitated with Lsm1–Lsm7, although the efficiency of co-immunoprecipitation varied (Fig. 3a). No co-immunoprecipitation of Dcp1 with Lsm8 could be detected, however, even after long exposures. Similarly, Pat1/Mrt1 also co-immunoprecipitates with Lsm1–Lsm7 proteins, but not with Lsm8 (Fig. 3b; and data not shown).

The Lsm and Sm proteins are RNA-binding complexes and function in splicing by assembling with snRNAs^{3,4}. To investigate whether Lsm proteins affect decapping through interactions with mRNA, we determined whether mRNAs co-immunoprecipitated with Lsm proteins (Fig. 4). Lsm1, Lsm5 and Dcp1 co-immunoprecipitated with a small (1–3% of the total) but clearly detectable amount of *CYH2* mRNA. In contrast, Lsm8 did not co-immunoprecipitate the *CYH2* mRNA in amounts significantly above background levels. Moreover, we did not detect U6 snRNA in immunoprecipitates of Lsm1 and Dcp1, which indicates that

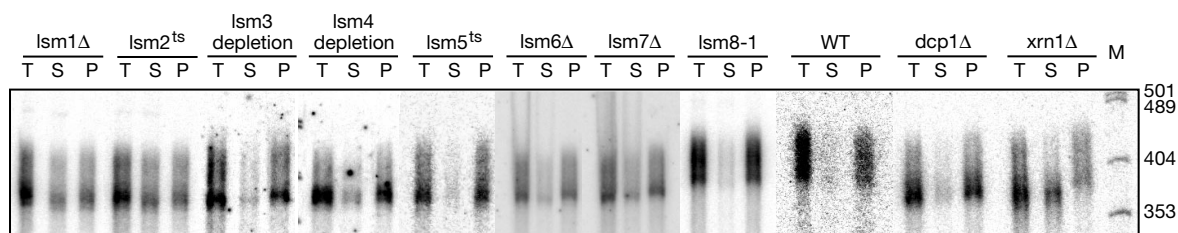


Figure 2 Lsm mutants accumulate oligoadenylated, full-length capped mRNAs. Steady-state mRNA from various *lsm* mutants, *dcp1Δ*, *xrn1Δ* and wild-type strains were analysed on a polyacrylamide northern blot with or without immunoprecipitation with anti-cap antibodies. Strains expressing *LSM3* and *LSM4* under *GAL* control were shifted to

glucose for 12 h before analysis, which has been shown to deplete these proteins⁷. T, total RNA; P, RNA extracted from the immunoprecipitate; S, supernatant. Radiolabelled *Hpal*-digested pUC18 DNA was used as marker (M).

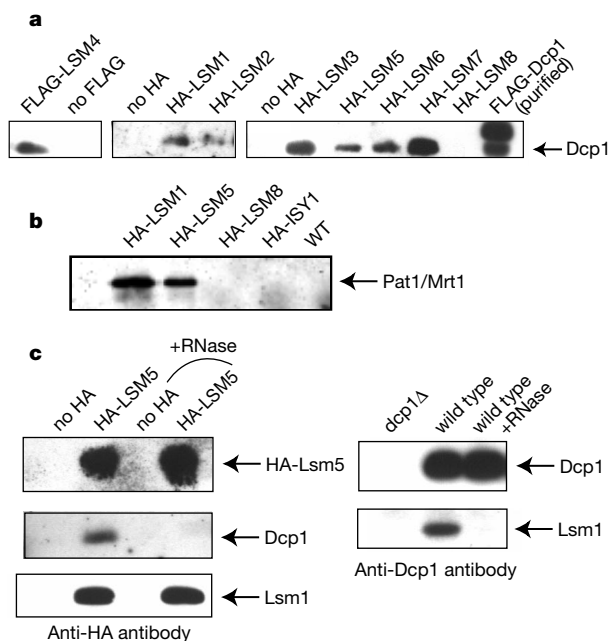


Figure 3 Lsm proteins co-immunoprecipitate with proteins involved in mRNA decay. **a**, Dcp1 co-immunoprecipitates with proteins Lsm1–Lsm7. Epitope-tagged Lsm proteins were immunoprecipitated and the immunoprecipitates were analysed by western blot for Dcp1. **b**, Co-immunoprecipitation of Pat1 with HA-tagged Lsm proteins. Epitope-tagged Lsm proteins were immunoprecipitated and the immunoprecipitates were analysed by western blot for Pat1. Isy1, a protein involved in pre-mRNA splicing²², was used as a

control. **c**, Interaction between Dcp1 and Lsm proteins is RNA dependent. Immunoprecipitation of Dcp1 (right panel) and HA–Lsm5 (left panel) from lysates of different strains (indicated above the lanes) was done with (+RNase) or without prior RNase treatment and presence of co-immunoprecipitating proteins (Dcp1 and Lsm1) was analysed by western blot using polyclonal Dcp1 or Lsm1 antisera.

these interactions are RNA specific (Fig. 4b; and data not shown). These results suggest that Lsm proteins interact with mRNA molecules and thereby modulate decapping. The co-immunoprecipitation of Lsm proteins with Dcp1, but not with each other or Pat1, is sensitive to RNase treatment (Fig. 3c and data not shown). This suggests that either the Lsm complex and Dcp1 interact directly in a manner dependent on one or both of these proteins binding mRNA, or that these proteins interact simultaneously with the same mRNA molecules. Thus, the co-immunoprecipitation of Lsm proteins, Dcp1 and mRNA may reflect the detection of a transient mRNP (ribonuclear protein particle) organization that exists before mRNA decapping.

These physical interactions suggest that the Lsm proteins have a role in cytoplasmic mRNA degradation. Consistent with this view, Lsm1, Lsm7, Dcp1 and Pat1 can be detected in the cytoplasm by immunofluorescence (Fig. 5). Lsm7 was also found in both compartments, consistent with it affecting both mRNA decay and splicing^{6,7}. Some nuclear staining was observed for both Dcp1 and Lsm1, suggesting that these proteins may also have a nuclear function.

Several observations indicate that a complex, consisting of Lsm1–Lsm7, is involved in mRNA decay and that a different complex, consisting of Lsm2–Lsm8, is associated with U6 and involved in pre-mRNA splicing. First, although mutations in *LSM2–LSM7* affect both mRNA decay and pre-mRNA splicing, mutations in *LSM1* only affect mRNA decay, and lesions in *LSM8* affect splicing but not mRNA decay (Fig. 1)^{7,8,10}. Second, proteins Lsm2–Lsm7 co-immunoprecipitate with both U6 snRNA^{6,7,9} and mRNA decay factors (Fig. 3), whereas Lsm1 does not immunoprecipitate U6 snRNA^{6,7} but does pull down Dcp1 and Pat1/Mrt1 (Fig. 3). In contrast, Lsm8 does co-immunoprecipitate U6 snRNA^{6–9} but does not pull down Dcp1 or Pat1/Mrt1 (Fig. 3). Third, immunoprecipitation of Lsm8 co-immunoprecipitates Lsm2–Lsm7, but not Lsm1 (ref. 6). Fourth, a pair of distinct complexes, each consisting of seven members, would be consistent with the known organization of Sm and Lsm proteins¹⁹. This suggests that the Lsm proteins form at least two distinct complexes that are involved in different aspects of RNA metabolism.

The function of the Lsm proteins in decapping appears to be to interact with the mRNA substrate and accelerate decapping, perhaps by promoting rearrangements in mRNP organization that

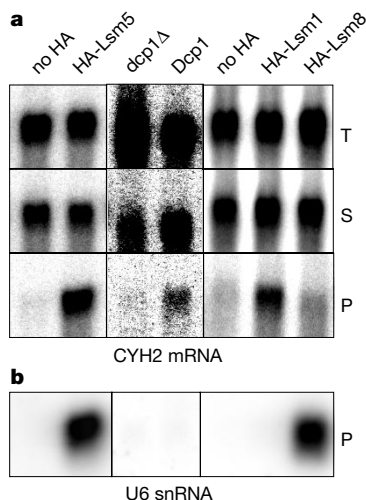


Figure 4 Lsm proteins and Dcp1 co-immunoprecipitate mRNA. RNA was extracted from the immunoprecipitate (P) of Dcp1, HA–Lsm5, HA–Lsm1 or HA–Lsm8 and analysed by northern blot for *CYH2* mRNA (a) or U6 snRNA (b). Similar analysis of aliquots of the lysate (T), and supernate (S) for the *CYH2* mRNA is also shown (a).

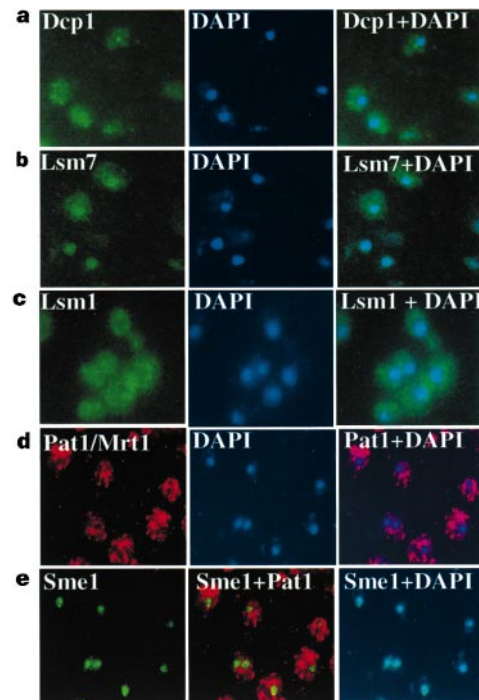


Figure 5 Cellular localization of Dcp1, Pat1/Mrt1, Lsm7 and Lsm1 by immunofluorescence analysis. **a–c**, Strains expressing HA-tagged Dcp1 (AEMY121 (a)), Lsm7 (AEMY51 (b)) or Lsm1 (AEMY28 (c)) were stained with anti-HA antibodies and anti-rat (fluorescein isothiocyanate) secondary antibody (green fluorescence, left panel), or subjected to 4,6-diamidino-2-phenylindole (DAPI) nuclear staining (blue fluorescence, middle panel). A merge of the two panels is shown in the right panel in each case. **d**, AEMY54 cells were stained with anti-Pat1 antibodies and anti-rabbit (cy3) secondary antibody (red fluorescence, left panel), or DAPI nuclear stain (blue fluorescence, middle panel). **e**, As a control, cells expressing the HA-tagged Sme1 nuclear protein (AEMY54) were stained with anti-HA antibodies and anti-rat (fluorescein isothiocyanate) secondary antibody (green fluorescence, left panel). A merge of HA–Sme1 staining and Pat1 staining (middle panel), and a merge of HA–Sme1 staining and DAPI nuclear staining (right panel).

lead to dissociation of the cytoplasmic cap-binding complex, which is an inhibitor of decapping²⁰. A general role for Lsm proteins in modulating RNP structure in a variety of contexts by affecting RNA–RNA and/or protein–RNA interactions would be consistent with the proposal that their function in pre-mRNA splicing is to facilitate RNP rearrangements during splicing^{7,21}. □

Methods

Yeast strains and plasmids

All strains used for decay analysis were in the genetic background of yRP841 (*MATα trp1 ura3-52 leu2-3,112 lys2-201 cup1Δ::LEU2(PM)*)¹², except the *lsm8-1* allele, which was compared with an isogenic wild-type control. Each strain and its genotypic difference from yRP841 are as follows: yRP1365 (*lsm1Δ::TRP1*); yRP1366 (*lsm2Δ::HIS3* pAEM55 (carrying a protein fusion that confers temperature sensitivity to Lsm2); yRP1367 (*his4-539 lsm5Δ::TRP1* pRP949 (carrying a protein fusion that confers temperature sensitivity to Lsm5); yRP1369 (*lsm6Δ::HIS3*); yRP1370 (*lsm7Δ::HIS3* LYS2); yRP1372 (*MATα his4-539 pat1Δ::LEU2*); yRP1407 (*MATα lsm1Δ::TRP1 lsm6Δ::HIS3*); yRP1070 (*dcp1Δ::URA3*)¹⁴; yRP884 (*xrn1Δ::URA3*)¹¹; BP4 (ref. 8) (*MATα lsm8-1*) and its isogenic wild-type control CY3 (ref. 8) (*MATα ura3 lys2 ade2 trp1 his3 leu2*). All strains used for immunoprecipitation, except FLAG–Lsm4 strain, were in the same genetic background, expressing an haemagglutinin A (HA)-tagged Lsm protein and compared with isogenic wild-type strains lacking any epitope tags (BMA64)⁷. Each strain and its genotypic difference from BMA64 (ref. 7) (*MATα trp1Δ1, his3-11, -15, ura3-1, leu2-3, -112, ade2-1, can-100*) are as follows: AEMY30 (*trp1-1 his3Δ200 lsm2Δ::HIS3* pAEM55 (HA–Lsm2); AEMY33 (ref. 7) (*trp1-1 his3Δ200 lsm2Δ::HIS3* pAEM55 (HA–Lsm2); AEMY33 (ref. 7) (*lsm3Δ::TRP1* pAEM64 (GAL–HA–Lsm3)); AEMY28 (ref. 7) (HA–Lsm1); AEMY45 (*MATα lsm8Δ::TRP1* pAEM76 (GAL–HA–Lsm8)); AEMY47 (ref. 7) (*lsm5Δ::TRP1* pAEM75 (GAL–HA–Lsm5)). AEMY49 (*MATα LSM6–HA–TRP1*), AEMY50 (*MATα*

LSM7-HA-TRP1); IDY2 (ref. 22) (*MATa trp1-1 his3Δ200 isy1Δ::HIS3 pCA23 (HA-ISY1, CEN URA3)*). For Lsm4, the MCY4 (ref. 5) strain expressing *GPD* promoter-controlled FLAG-Lsm4 protein from a plasmid (pRP961) was used. This strain has a chromosomal *LSM4* gene under *GAL* UAS control and was grown in glucose medium for the experiment. MCY4 (ref. 5) (*MATa LEU2-GAL-LSM4*) was compared with its isogenic wild-type strain w303 (*MATα ade1-101 his3-Δ1 trp1-289 ura33-52*) without epitope-tagged proteins.

RNA analysis

Transcriptional shut-off experiments, northern analysis and immunoprecipitation of RNA were performed as described^{12,13,15}. Oligo oRP140 (5'-ATATTGATTAGATCAGG-AATTCC-3') was used to probe for *MFA2pG* mRNA.

Immunology

The immunoprecipitation and immunofluorescence experiments followed standard protocols. The HA-Lsm protein immunoprecipitations shown in Fig. 3a and c were done using anti-HA antibody sepharose. Lysis and washing buffer was IPD1 buffer (50 mM Tris pH7.5, 50 mM NaCl, 2 mM MgCl₂, 0.1% Nonidet P-40) containing 1 mM β-mercaptoethanol and 1× 'COMPLETE' protease inhibitor mix (Boehringer Mannheim). Bound proteins were eluted with 50 mM Tris pH7.5, 100 mM NaCl, 2 mM MgCl₂, 1% SDS, 1 mg ml⁻¹ HA peptide and 1× COMPLETE protease inhibitor mix, and analysed for presence of Dcp1 by western blotting²³. The Dcp1 immunoprecipitation shown in Fig. 3c was done similarly using anti-Dcp1 polyclonal antibodies and protein A sepharose, and the co-immunoprecipitating Lsm1 was revealed in western blots using anti-Lsm1 polyclonal antibodies. Elution buffer was as above but without HA peptide. The immunoprecipitation of FLAG-Lsm4 shown in Fig. 3a was done similarly using anti-FLAG antibody agarose (Sigma). Lysis was in IPD2 buffer (50 mM Tris pH7.5, 25 mM NaCl, 2 mM MgCl₂, 0.2% Nonidet P-40). IPD2 buffer without Nonidet P-40 was used to wash the beads. Elution buffer was as above but with 200 μg ml⁻¹ FLAG peptide. HA-Lsm immunoprecipitation in Fig. 3b was done using anti-HA antibodies and protein A sepharose. Lysis and washing buffer was 6 mM HEPES pH7.9, 150 mM NaCl, 2.5 mM MgCl₂, 0.05% Nonidet P-40. Immunoprecipitated proteins were extracted by boiling pellets in SDS sample buffer and western analysed using anti-Pat1 polyclonal antibodies. To study the RNA dependence of co-immunoprecipitations, lysates were treated with 0.5 μg μl⁻¹ RNase A for 10 min at room temperature before preclearing. To study the co-immunoprecipitation of RNA, total RNA extracted from the immunoprecipitates was analysed by northern blotting using *CYH2* complementary DNA probe or U6 oligo probe (oRP 759: 5'-GACCAATGTCACGAAGG-3').

Received 29 July 1999; accepted 10 February 2000.

1. Beelman, C. A. & Parker, R. Degradation of mRNA in eukaryotes. *Cell* **81**, 179–183 (1995).
2. Jacobson, A. & Peltz, S. W. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* **65**, 693–739 (1996).
3. Seraphin, B. Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.* **14**, 2089–2098 (1995).
4. Hermann, H. et al. snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. *EMBO J.* **14**, 2076–2088 (1995).
5. Cooper, M., Johnston, L. H. & Beggs, J. D. Identification and characterization of Uss1p (Sdb23p): a novel U6 snRNA-associated protein with significant similarity to core proteins of small nuclear ribonucleoproteins. *EMBO J.* **14**, 2066–2075 (1995).
6. Salgado-Garrido, R., Bragado-Nilsson, E., Kandels-Lewis, S. & Seraphin, B. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J.* **18**, 3451–3462 (1999).
7. Mayes, A. E., Verdono, L., Legrain, P. & Beggs, J. D. Characterization of Sm-like proteins in yeast and their association with U6 snRNA. *EMBO J.* **18**, 4321–4331 (1999).
8. Pannone, B. K., Xue, D. & Wolin, S. L. A role for the yeast La protein in U6 snRNP assembly: evidence that the La protein is a molecular chaperone for RNA polymerase III transcripts. *EMBO J.* **17**, 7442–7453 (1998).
9. Stevens, S. W. & Abelson, J. Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins. *Proc. Natl Acad. Sci. USA* **96**, 7226–7231 (1999).
10. Boeck, R., Lapeyre, P., Brown, C. E. & Sachs, A. B. Capped mRNA degradation intermediates accumulate in the yeast *spb8-2* mutant. *Mol. Cell. Biol.* **18**, 5062–5072 (1998).
11. Decker, C. J. & Parker, R. A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation. *Genes Dev.* **7**, 1632–1643 (1993).
12. Hatfield, L., Beelman, C. A., Stevens, A. & Parker, R. Mutations in trans-acting factors that inhibit mRNA decapping in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **16**, 5830–5838 (1996).
13. He, W. & Parker, R. Analysis of mRNA decay pathway in *S. cerevisiae*. *Methods* **17**, 3–10 (1999).
14. Beelman, C. A. et al. An essential component of the decapping enzyme required for normal rates of mRNA decay in yeast. *Nature* **382**, 642–646 (1996).
15. Dunkley, T. & Parker, R. The DCP2 protein is required for mRNA decapping in *Saccharomyces cerevisiae* and contains a functional MutT motif. *EMBO J.* **18**, 5411–5422 (1999).
16. Muhrad, D., Decker, C. J. & Parker, R. Deadenylation of the unstable mRNA encoded by the yeast *MFA2* gene leads to decapping followed by 5' → 3' digestion of the transcript. *Genes Dev.* **8**, 855–866 (1994).
17. Fromont-Racine, M. et al. Interactions of the Lsm proteins of *S. cerevisiae*. *Proc. Natl Acad. Sci. USA* (submitted).
18. Wang, X., Watt, P. M., Louis, E. J., Borts, R. H. & Hickson, I. D. Pat1: a topoisomerase II-associated protein required for faithful chromosome transmission in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **24**, 4791–4797 (1996).
19. Kambach, C. et al. Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387 (1999).
20. Schwartz, D. & Parker, R. Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **19**, 5247–5256 (1999).

21. Achsel, T. et al. A doughnut-shaped heteromer of human Sm-like proteins binds to the 3'-end of U6 snRNA, thereby facilitating U4/U6 duplex formation *in vitro*. *EMBO J.* **18**, 5789–5802 (1999).
22. Dix, I., Russell, C., Yehuda, S. B., Kupiec, M. & Beggs, J. D. The identification and characterization of a novel splicing protein, Isy1p, of *Saccharomyces cerevisiae*. *RNA* **5**, 360–368 (1999).
23. Tharun, S. & Parker, R. Analysis of mutations in the yeast mRNA decapping enzyme. *Genetics* **151**, 1273–1285 (1999).

Acknowledgements

We thank members of R.P.'s laboratory for useful discussions and comments on the manuscript. This study was supported by funds from the Howard Hughes Medical Institute, the NIH, a Wellcome Trust Prize Studentship to A.E.M. and a Royal Society Cephalosporin Fund Senior Research Fellowship to J.D.B. We also thank N. Rodriguez-Cousino for the antisera to Pat1, C. Russell for providing the HA-Isy1 strain and A. Sachs for the anti-Lsm1 antibodies.

Correspondence and requests for materials should be addressed to R.P. (e-mail: rrparker@u.arizona.edu).

Structural basis for the anticoagulant activity of the thrombin–thrombomodulin complex

Pablo Fuentes-Prior*, Yoriko Iwanaga*, Robert Huber*, Rene Pagila†, Galina Rumennik†, Marian Seto†, John Morser†, David R. Light† & Wolfram Bode*

* Max-Planck-Institut für Biochemie, Abteilung Strukturforschung Am Klopferspitz 18a, D-82152 Martinsried, Germany
† Berlex Biosciences, Richmond, California 94804, USA

The serine proteinase α-thrombin causes blood clotting through proteolytic cleavage of fibrinogen and protease-activated receptors and amplifies its own generation by activating the essential clotting factors V and VIII¹. Thrombomodulin², a transmembrane thrombin receptor with six contiguous epidermal growth factor-like domains (TME1–6), profoundly alters the substrate specificity of thrombin from pro- to anticoagulant by activating protein C (see, for example, reference 2). Activated protein C then deactivates the coagulation cascade by degrading activated factors V and VIII². The thrombin–thrombomodulin complex inhibits fibrinolysis by activating the procarboxypeptidase thrombin-activatable fibrinolysis inhibitor³. Here we present the 2.3 Å crystal structure of human α-thrombin bound to the smallest thrombomodulin fragment required for full protein-C co-factor activity, TME456. The Y-shaped thrombomodulin fragment binds to thrombin's anion-binding exosite-I, preventing binding of procoagulant substrates. Thrombomodulin binding does not seem to induce marked allosteric structural rearrangements at the thrombin active site. Rather, docking of a protein C model to thrombin–TME456 indicates that TME45 may bind substrates in such a manner that their zymogen-activation cleavage sites are presented optimally to the unaltered thrombin active site.

Thrombomodulin consists of a lectin-like amino-terminal domain, followed by a hydrophobic segment, six contiguous epidermal growth factor (EGF)-like domains, an O-glycosylated serine/threonine-rich domain, a transmembrane segment and a short cytoplasmic tail (Fig. 1a). Upon thrombomodulin binding, thrombin activation of both protein C and thrombin-activatable fibrinolysis inhibitor (TAFI) is enhanced by at least three orders of magnitude^{2,3}. Different hypotheses have been proposed to explain the thrombomodulin-induced switch in thrombin specificity. Mutagenesis studies (see, for example, ref. 4) and experiments with fluorescence⁵ or spin label⁶ active-site probes indicate that thrombomodulin binding may cause allosteric changes in the

environment of the thrombin active site. However, calorimetric and enzyme kinetic experiments⁷ show that thrombomodulin binding is coupled only to those allosteric changes in the thrombin active site that result from saturation of the sodium-binding site⁸. The activity of free thrombin, even if saturated by sodium ions, is not adequate to meet the physiological requirements for activation of protein C and TAFI. Therefore, to investigate how thrombomodulin analogues containing the critical TME4 domain enhance activation of protein C by thrombin, we solved the crystal structure of the stoichiometric complex between human α -thrombin and TME456 at 2.3 Å resolution.

The four independent complexes of L-Glu-Gly-Arg chloromethyl ketone (EGR-CMK)-inhibited thrombin and TME456 have almost identical structures, which are thus unaffected by crystal contacts. Each consists of an ellipsoidal α -thrombin molecule and an angular Y-shaped TME456 fragment (Fig. 1b). TME4 and much of TME6 protrude from the thrombin surface, whereas TME5 and part of TME6 bind to a cluster of lysine and arginine residues distant from the thrombin active site (termed anion-binding exosite-I⁹), in agreement with previous evidence^{10–12}. Macromolecular substrates and inhibitors required an intact anion-binding exosite-I for thrombin recognition and complex formation⁹. Occupation of this region by TME56 prevents binding of procoagulant substrates (for example, fibrinogen or protease-activated receptors), impairing their cleavage by thrombomodulin-bound thrombin.

In the thrombomodulin–thrombin complex, the main chain of thrombin departs only slightly from other reported thrombin structures. This point is illustrated in Fig. 2 by comparing thrombin in the thrombomodulin complex with thrombin bound to the exosite-I-binding inhibitor triabin¹³ (Fig. 2a) or with the prototypical D-Phe-Pro-Arg chloromethyl ketone (PPACK)-inhibited thrombin structure⁹ (Fig. 2b). This high degree of structural con-

servation also applies to the active-site residues of thrombin and the surrounding acidic residues Glu 217, Glu 192 and Glu 39 (using the chymotrypsinogen numbering of thrombin⁹), which have been identified as possible candidates to explain the thrombomodulin-mediated conversion of thrombin to an anticoagulant proteinase (see, for example, ref. 4). Displacements are restricted to the prominent 60-loop, whose exposed Tyr 60A–Trp 60D bulge pulls up to 1.7 Å away from the active site compared to PPACK-thrombin. The nearby 37- and 96-loops diverge even less, whereas the 149-loop is disordered in thrombomodulin-bound thrombin, as frequently observed.

These conformational differences are comparable to those reported in other thrombin structures in complex with various active site or exosite-I-binding inhibitors, and result mainly from crystal packing. The virtually unaltered active-site conformation of thrombin–thrombomodulin indicates that allosteric modification of the thrombin active site is not the main mechanism by which thrombomodulin enhances activation of protein C. This is supported by the finding that TME56 is the only thrombomodulin substructure that interacts with thrombin in chondroitin-sulphate-free thrombomodulin (Fig. 1b), but does not increase activation of protein C when it alone binds to thrombin¹⁰. TME4, the domain that is responsible for modulating the substrate specificity¹⁴, protrudes away from the proteinase (Fig. 1b). These observations are consistent with the finding that peptide substrates are cleaved with similar kinetics by thrombin bound to thrombomodulin or the exosite-I-binding peptide hirugen⁷. Alterations in the thrombin active site that were detected upon thrombomodulin binding^{5,6} may therefore reflect changes in the electrostatic environment of the proteinase and/or small structural rearrangements.

The fundamental change in thrombin specificity appears to reside in the three-dimensional arrangement of EGF-like repeats observed

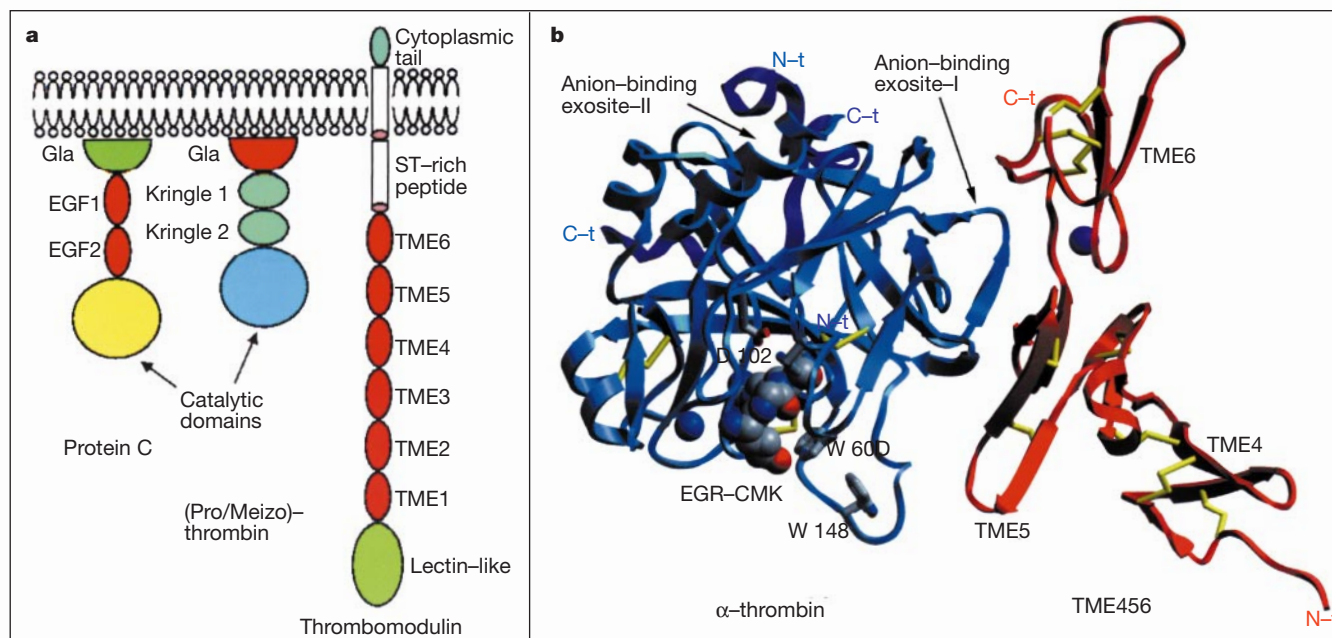


Figure 1 The protein C anticoagulant pathway. **a**, Domain organization of protein C, prothrombin and thrombomodulin. Protein C/activated protein C, prothrombin and the active precursor meizothrombin bind to phospholipid membranes through their highly homologous γ -carboxyglutamate (Gla)-rich domains. Des(F1)meizothrombin²⁶ and α -thrombin⁹ lack the Gla-domain and kringle 1, or both kringle domains, respectively. **b**, Ribbon model of the complex between α -thrombin (pink, light chain; blue, heavy chain) and TME456 (red). The irreversible inhibitor L-Glu-Gly-Arg chloromethyl ketone (EGR-

CMK) mimics the acidic P3 residue of the protein C activation peptide and is shown as a space-filling model with atoms colour-coded (carbon, grey; nitrogen, blue; oxygen, red). The disulphide linkages are shown in yellow. The N and C termini of all three chains are labelled. The side chains of active site residues His 57 and Asp 102 as well as Trp 60D and Trp 148 are explicitly shown. The buried calcium ion is shown as a pink sphere, and the Na⁺ ion as a blue sphere.

in TME456 (Figs 1, 3). EGF-like domains often occur in multiple copies, forming extended structures^{15,16} (Fig. 3a). A rod-like arrangement of the three EGF-like domains in TME456 was anticipated¹⁷, based on the sequence similarity between thrombomodulin and fibrillin-1 (ref. 16). Instead, TME4 is anchored nearly perpendicular to the linear TME56 tandem (Figs 1, 3).

The TME4–5 connecting segment (Gln 387–Phe 389) is clamped to TME4 by hydrogen bonds and to TME5 through major hydrophobic contacts. In particular, Met 388 slots into a hydrophobic groove of TME5 spanned by the disulphide bridges [1–2] and [5–6] to anchor TME4 to TME5 and to direct TME4 away from thrombin (Fig. 3b). Oxidation of Met 388 (ref. 18) and all insertions, deletions and most amino-acid substitutions in this segment¹⁹ drastically reduce anticoagulant activity, underlining the importance of the proper orientation between TME4 and the TME56 tandem.

At the TME5–6 interface, the conserved internal side chain of Tyr 413 (TME5) makes strong contacts with the β 2– β 3 loop of TME6 (Fig. 3b). Highlighting this essential structural role, the Tyr413Ala mutant lacks co-factor activity²⁰. A similar role is played in fibrillin-1 by Tyr 2157 (ref. 16), but the bending of TME5 allows additional interdomain contacts in TME456 (for example, Asn 391–Pro 441). These contacts may further protect the bound Ca^{2+} ion in TME6 from bulk solvent, explaining its

markedly low dissociation constant (K_d)²¹.

Overall, TME4 and TME6 fold similarly to other EGF-like domains¹⁵ and consist of a flat major and a twisted minor β -sheet with the usual [1–3, 2–4, 5–6] disulphide pairing. However, they differ considerably in detail (Figs 1b, 3, 4). We note that residues Val 371–Phe 389 adopt a conformation in solution that corresponds closely to its structure within TME456, consistent with our proposal that this peptide constitutes a nucleation site for folding²². TME6 also follows the classical scheme; however, the short stretch that corresponds to the minor sheet (residues Gly 449–Ala 455, disordered in all four TME456 molecules) precedes an additional strand that aligns antiparallel to β 3 (Figs 3, 4). Cardinal to the overall domain folding is a calcium ion with octahedral coordination (Fig. 4) predicted by mutagenesis²⁰ and calcium-binding studies²¹. The central TME5 domain diverges strongly from archetypal EGF structures because of its unusual [1–2, 3–4, 5–6] disulphide pairing, detected in a previous nuclear magnetic resonance (NMR) investigation²³. Its extended major sheet (strands β 2 and β 3 and the connecting loop) projects into thrombin's active site groove (Fig. 1). This projection has a well defined electron density, although the NMR analysis²³ suggests some flexibility (Fig. 3a).

An interface of $\sim 900 \text{ \AA}^2$ is buried between TME56 and the anion-binding exosite-I of thrombin. Electrostatic interactions of the

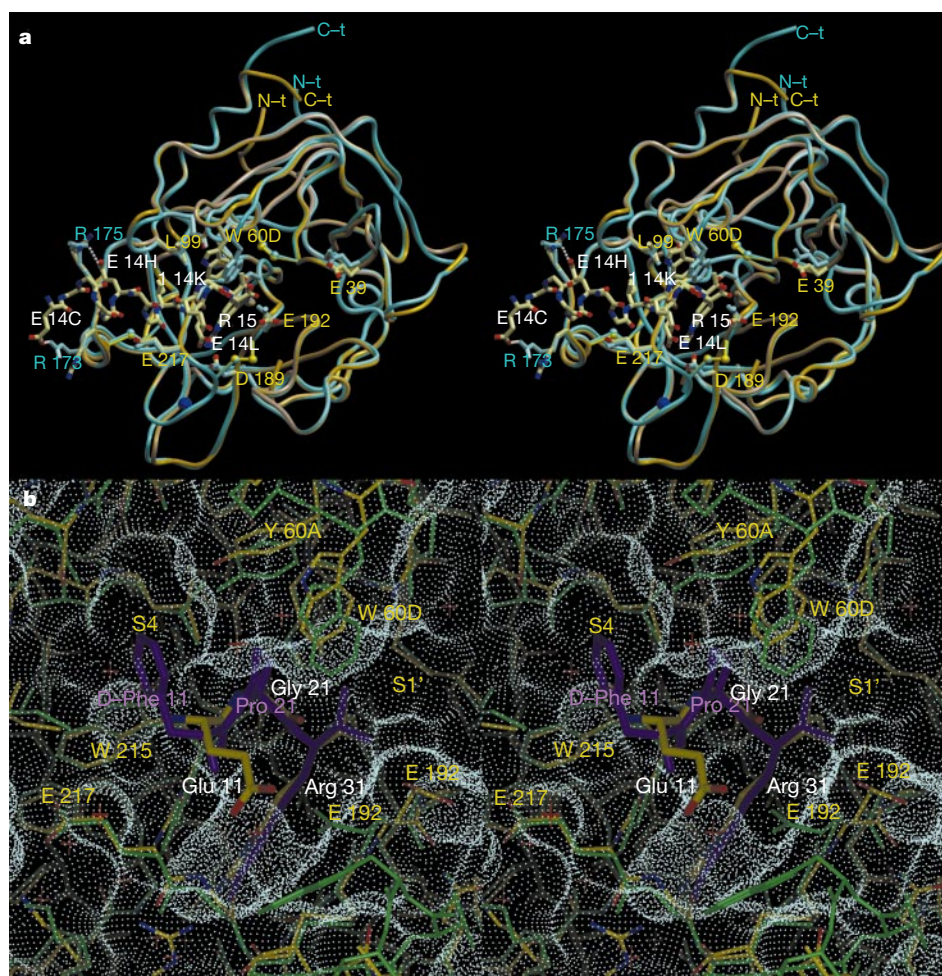


Figure 2 Thrombomodulin is not an allosteric modulator of thrombin. **a**, Stereo view of thrombomodulin-bound thrombin (yellow, side chains colour-coded: carbon, yellow; nitrogen, blue; oxygen, red; Na^+ ion depicted as a blue sphere), overlaid with triabin-bound thrombin (light blue). The light chain of a neighbouring thrombin molecule binds in a substrate-like manner in the active site of thrombin–triabin and is shown colour-coded.

b, Close-up of the active site region (colour-coded; Connolly surface shown as blue dots), overlaid on the structure of PPACK–thrombin (green). Only the immediate surrounding of active-site residues and the chloromethyl ketone moieties (EGR-CMK, colour-coded; PPACK purple) are represented. In both cases, the r.m.s. deviations are below $0.60 \text{ \AA}$ for $>280 \text{ C}\alpha$ pairs.

oppositely charged moieties (Fig. 5a,b) certainly guide formation of the tight thrombin–thrombomodulin complex ($K_d \approx 4$ nM)¹⁹. However, most of the basic residues of thrombin either form solvent-mediated hydrogen bonds with uncharged groups on thrombomodulin or extend freely into the bulk solvent (Fig. 5c, d). The preponderance of hydrophobic contacts is exemplified by Ile 414, the alkyl side chain of which fits in a surface depression on thrombin formed by Phe 34, Tyr 76 and Ile 82. This hydrophobic patch is completed by the side chains of Leu 65 (thrombin) and Ile 424 (thrombomodulin) (Fig. 5d). Alanine mutations at Ile 414

and Ile 424 lead to greatly reduced thrombin affinities and co-factor activities²⁰, emphasizing the importance of both residues. The bound calcium ion stabilizes the TME56 junction into a shape complementary to thrombin's exosite I and increases the affinity of soluble thrombomodulin fragments by two orders of magnitude²¹. This feature explains the greatly reduced activity of alanine replacements for calcium ligands Asp 423, Glu 426 and Asn 439 (ref. 20).

The structure of thrombin complexed with the reduced non-adeuropeptide Glu 408–Glu 426 from TME5 has been reported¹¹.

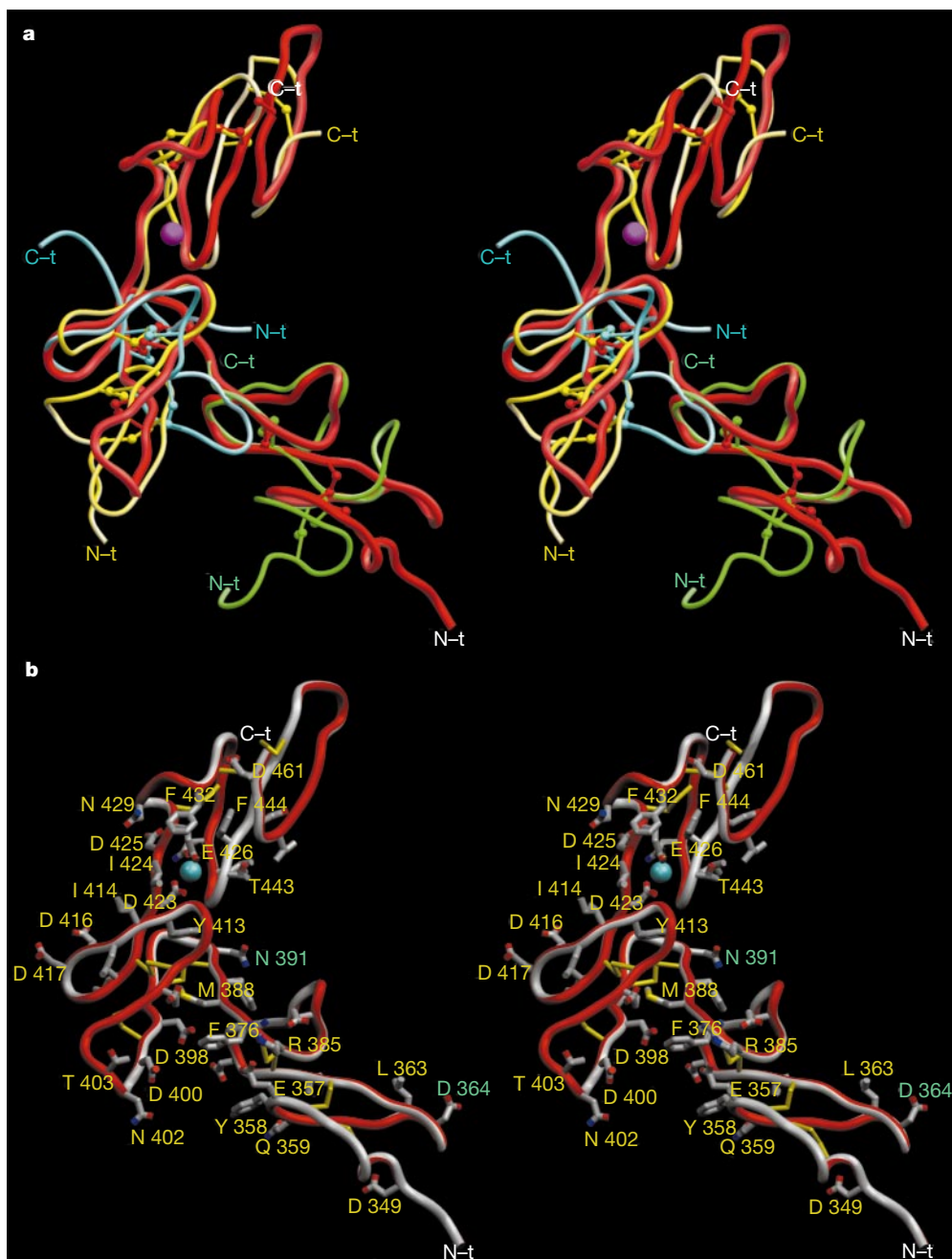


Figure 3 TME456 exhibits a novel arrangement of EGF-like repeats. **a**, Ribbon plot of TME456 (red) overlaid with the lowest energy solution structures of isolated TME4 (green), TME5 (light blue) and the pair of EGF-like repeats of fibrillin-1 (yellow). Disulphide bridges and the N and C termini of each structure are indicated. Note the differences between the structures of EGF domains TME4 and TME5 within TME456 and the corresponding conformations of the isolated domains in solution^{23,34}. This compact structure of TME456

accounts for the small nodule observed in electron micrographs of soluble thrombomodulin, whereas the opposite lectin-like domain is visualized as a larger globular nodule³⁵. **b**, Ribbon plot, highlighting side chains of residues critical for co-factor activity. Side chains N-glycosylated in the native protein, Asn 364 (here Asp, after PNGase treatment) and Asn 391 are also shown (labelled green).

This peptide sandwiches between the exosite-I regions of two α -thrombin molecules in the reported crystal structure. The path followed by the peptide while contacting one of the thrombin molecules corresponds closely to the conformation of residues Cys 407–Asp 425 in the thrombin–TME456 complex. However, most peptide residues are shifted by one compared to the corresponding segment in the thrombin–TME456 structure, with unsatisfactory consequences. For example, in their structure the side chain of Tyr 413 contacts thrombin (instead of the above mentioned Ile 414), and Cys 409 and Cys 421 are not in a position to form a disulphide bond.

There is a notable asymmetry in the electrostatic surface potential of TME456. The thrombin-binding surfaces of TME5 and TME6 are more negative than the surface facing away from the proteinase (Fig. 5b). The outer TME4 surface is noticeably hydrophobic (Leu 363, Leu 369, Pro 380). It harbours the fully glycosylated side chain of Asn 364, while the partially glycosylated Asn 391 emerges from the disulphide-closed loop β 1– β 2 of TME5 extending into the narrow outer groove between TME4 and TME5 (Fig. 3b). Carbohydrates attached to these asparagines are required neither for thrombin binding nor for co-factor activity, as shown by expression of active thrombomodulin in *Escherichia coli*^{20,22}. No residues relevant to thrombomodulin function map to this surface (Fig. 3b).

In contrast, the TME4 surface extending towards the active site groove carries a number of polar and charged side chains (Arg 385, His 384, Glu 382, Glu 357, Asn 355 and, more distally, Asp 349 and Glu 346) as well as several exposed aromatic residues (Phe 352, Tyr 358, Phe 376). The side chains of Asp 398, Asn 402, Thr 403, Gln 404 and Glu 408 of the projecting β 2– β 3 loop of TME5 point towards thrombin's active site. Most of these residues are indispensable for co-factor activity (see, for example, refs 20 and 22) (compare Fig. 3b). This clustering of functionally critical residues along an extended solvent-exposed region of TME45 strongly suggests that they are involved in substrate binding. We postulate that this surface represents the primary interacting region of the anticoagulant substrate, protein C.

To explore this possibility in more detail, we docked a protein C model based on the crystal structure of Gla-domainless activated protein C²⁴ (see Methods) to the thrombin–TME456 structure (Fig. 6). Insertion of the protein C model in the active-site groove of thrombin in a substrate-like manner (see Methods and Fig. 2a) satisfies important specificity requirements of thrombin. For instance, Arg 15, Pro 14 and Val 12 fill the S1, S2 and S4 subsites of the proteinase, and residues Glu 9 (P7) and Asp 4 (P12) form salt bridges with thrombin residues Arg 175 and Arg 173, respectively (Fig. 2a). Without additional assumptions, the docking experiment predicts important interactions between the activation peptide and the basic linker connecting the EGF domains of protein C, as well as between protein C and thrombin (Fig. 6). An intact Gla-EGF region is required for the binding of protein C to the thrombin–thrombomodulin complex²⁵. In our model, the EGF2 domain of the bound protein C makes favourable contacts with the kringle 2 domain of meizothrombin²⁶ (Figs 1a, 6), whereas the EGF1 domain is slightly shifted from the activated protein C crystal conformation²⁴ and binds in the groove between kringle 2 and the intermediate helix of (meizo)thrombin. The hydrophobic N-terminal loop of the preceding Gla-domain associates with the phospholipid cell membrane (Fig. 6).

Most relevant for the mechanism of protein C activation by thrombin–thrombomodulin, our model anticipates manifold interactions between the neighbouring 37- and 70-loops of protein C (chymotrypsinogen numbering²⁴) and TME45 (Fig. 6). The highly conserved Lys 37–Lys 38–Lys 39 triplet, which is important for activation of protein C catalysed by thrombin–thrombomodulin²⁷, projects towards the β 4– β 5 loop of TME4 and contacts Glu 382. The Ca^{2+} -binding 70–80 loop of protein C interacts with the β 1– β 2 loop of TME4 and the β 2– β 3 extension of

TME5; the critical protein C side chain of Arg 74 (ref. 28) is clamped between the carboxylates of Asp 398 and Glu 357 (Fig. 6). Evidence supporting the relevance of this electrostatic 'grip' includes: first, Asp398Ala and Glu357Ala²⁰ or Glu357Gln mutants²² are inactive, whereas the Glu357Asp mutant retains half of the wild-type co-factor activity²²; and second, the Arg74Gln variant of protein C yields fully active activated protein C when activated artificially, but resists activation by the thrombin–thrombomodulin complex²⁸.

Furthermore, protein C's Arg 74 guanidinium group is partially sandwiched between the critical aromatic side chains of Tyr 358 and Phe 376 (ref. 20) (TME4), contacting a solvent molecule of high electron density. We tentatively interpret this solvent molecule as a Na^+ ion present in the crystallization solution. However, it is conceivable that under physiological conditions this position is occupied by calcium. In fact, a Ca^{2+} bridge between thrombomodulin and protein C has been proposed¹⁴. Our model also predicts direct contacts between TME4 side chains and protein C residues that coordinate calcium (Arg 385–Glu 80, Fig. 6). Such interactions might explain the unique calcium dependency of the thrombin–thrombomodulin-catalysed activation of protein C^{2,14}.

The acidic TME3–4 linker may move towards the bound substrate, as suggested by the NMR structure of isolated TME4 (Fig. 3a), allowing residues Lys 62/Lys 63 (protein C) to interact with the side chain of Asp 349, critical for thrombin–thrombomodulin catalysed protein C activation^{14,20}. The Lys 62/Lys 63 pair is important as the

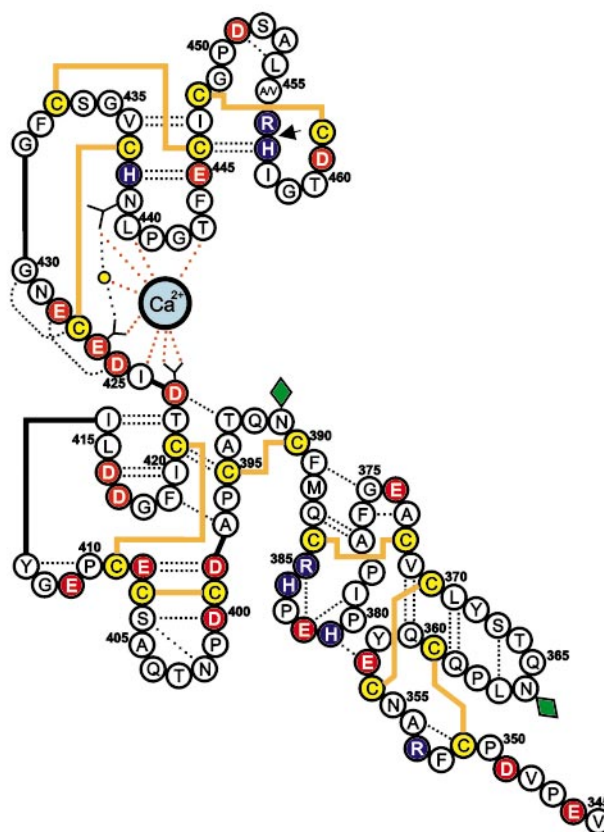


Figure 4 Sequence and secondary structure of the human TME456 fragment. Acidic residues are shadowed red, basic residues blue. Protein–protein hydrogen bonds are indicated with black dotted lines, those made with the bound calcium ion with red dotted lines. Disulphide bridges are indicated with solid orange lines, glycosylation sites with green diamonds. The trypsin cleavage site Arg 456–His 457 (shown with an arrow) has been replaced in the form used in this study by the Gly–Gln sequence conserved in thrombomodulin from other species to prevent proteolysis. Note that domains TME4 and TME5 are separated by an unusually short three-residue linker.

protein C double mutant [Lys62Ser, Lys63Asp] shows a 10-fold lower activation rate by thrombin–thrombomodulin¹⁷. This loop displacement would consequently bring the highly acidic TME3 domain (for example, Asp 338, Asp 341, Glu 343) close to the bound substrates and further stabilize the ternary complex, which might be essential for TAFI binding and processing.

These manifold, mainly electrostatic interactions would allow TME45 to align the substrate protein C presenting its scissile peptide bond Arg 15–Ile 16 to the thrombin active site residues with an optimal conformation for cleavage. Our model predicts a distance of ~50 Å between the scissile bond of protein C and the plane of the cell surface, slightly lower than the measured distance of 66 ± 5 Å (ref. 2). We are tempted to speculate that the extended binding surface formed by domains TME4 and TME5 similarly mediates additional physiological processes, such as TAFI activation and the inhibition of thrombomodulin-bound thrombin by the protein C inhibitor.

The current structure emphasizes the crucial role of secondary enzyme–substrate interactions for processing macromolecular substrates. Procoagulant substrates with non-optimal cleavage sites are efficiently activated upon binding to thrombin's exosite-I. Occupancy of exosite-I by TME56 blocks procoagulant activities by preventing binding of fibrinogen, PARs and co-factors V and VIII, without inducing major conformational changes in the active centre of thrombin. In the presence of calcium, protein C has a non-optimal cleavage site², and secondary contacts between its light chain and thrombin²⁵ are insufficient to overcome electrostatic

repulsion by the active site region of the proteinase. Thrombin binding to the acidic TME56 presents an additional substrate-binding surface on TME45. Thrombomodulin functions by promoting association of the scissile peptide bond of protein C/TAFI with the catalytic machinery of thrombin in a stereochemical conformation optimal for cleavage, manifested in the strong enhancement of the maximal cleavage rate (k_{cat}) and relatively small but favourable effects on the Michaelis constant (K_m) for cleavage¹⁹. Allosteric induction by TME4 cannot play an essential role in the thrombomodulin-mediated cleavage of protein C by thrombin. In this respect, thrombomodulin resembles the bacterial plasminogen activator staphylokinase, which also functions by presenting the non-optimal cleavage segment of plasminogen to the active site of plasmin²⁹. It is tempting to speculate that the clotting co-factors Va and VIIIa bring the activating proteinase (factors Xa, IXa) and the zymogen (prothrombin, factor X) together in a mode optimal for activation. □

Methods

Preparation of proteins

The Val 345–Cys 462 fragment of human thrombomodulin (with residues Arg 456 and His 457 mutated to Gly and Gln and the N terminus extended by a PhePro dipeptide) was expressed in Chinese hamster ovary cells. After anion-exchange chromatography¹⁹, the minor glycoform N-glycosylated at Asn 364 was further purified on a C4 column using an acetonitrile gradient in 0.1% TFA. This fraction was treated with recombinant N-glycosidase F (Boehringer Mannheim) and the carbohydrate-free Asp 364 analogue (TME456) was repurified by RP-HPLC and lyophilized as the TFA-salt. Human α -thrombin was isolated from frozen plasma according to standard procedures and irreversibly inhibited with an

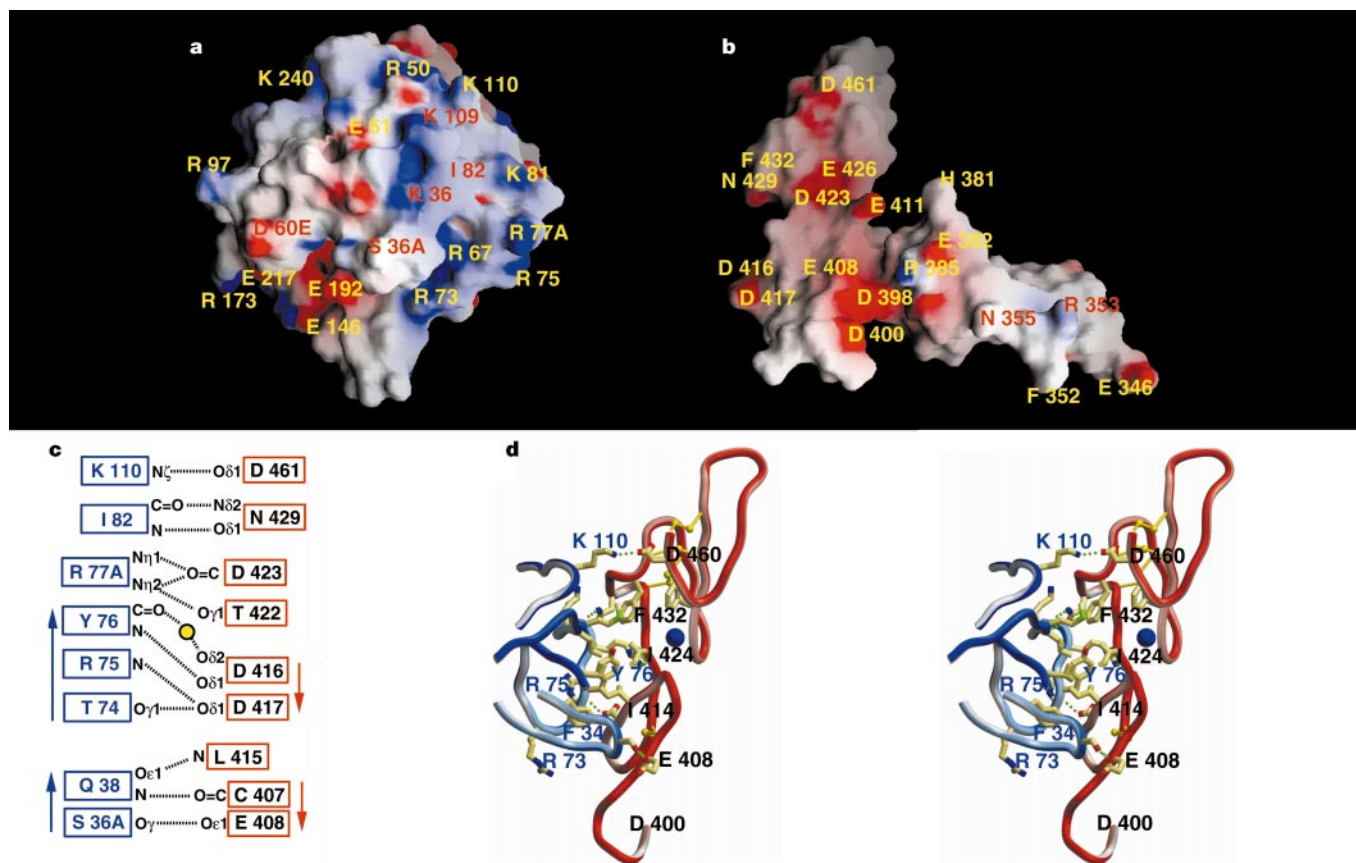


Figure 5 The thrombin–thrombomodulin interaction interface is dominated by hydrophobic contacts. GRASP³⁶ electrostatic surface potentials of α -thrombin (a) and TME456 (b). Both moieties have been slightly rotated around the y -axis to present their interaction interfaces to the viewer. Note the overall complementarity of electrostatic potentials. c, Major hydrogen-bonding interactions (dotted lines) between thrombin and

thrombomodulin. d, Detail of the interaction interface, highlighting its primary hydrophobic character. Selected hydrogen bonds are indicated as green spheres. Note that a single, solvent-exposed salt bridge forms between thrombin and thrombomodulin (Lys 110 N ζ –Asp 461 O δ 2). For clarity, only residues Asp 400–Cys 462 of thrombomodulin and the thrombin loops of the contact interface are shown.

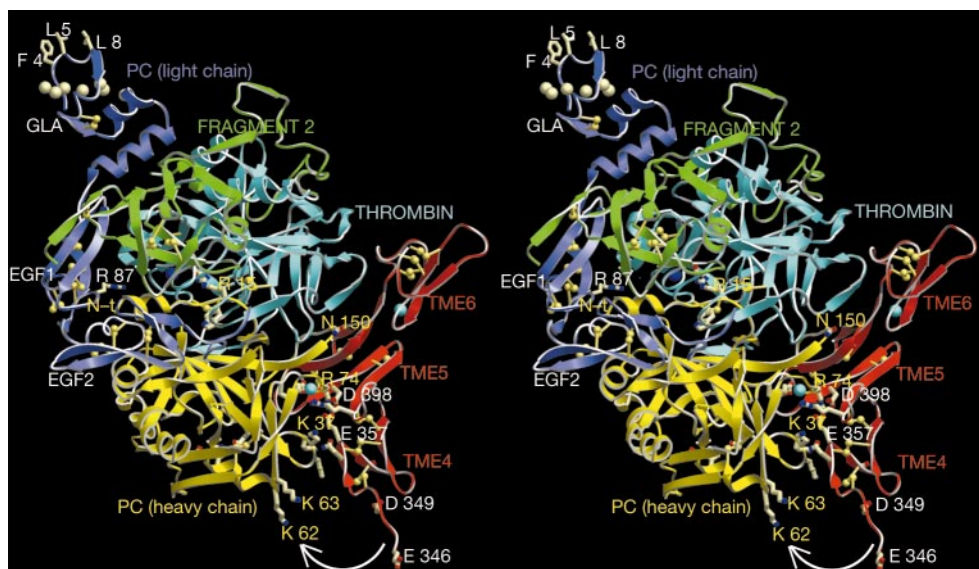


Figure 6 Hypothetical structure of the thrombin-thrombomodulin-protein C complex. The acidic N terminus of the activation peptide of protein C inserts between both EGF-like domains of protein C (PC) allowing intramolecular interactions with basic/polar residues (Arg 87, Gln 90, Arg 91, protein C light chain). The carboxylamide of Gln 118 and Lys 20 N ζ (protein C heavy chain, chymotrypsinogen numbering) satisfy the hydrogen-bonding potential of thrombin's Glu 192 carboxylate. Several hydrophobic side chains of protein C, together with the aliphatic parts of Lys 20 and Lys 159, form a shallow hydrophobic surface depression that accommodates Trp 60D, in agreement with other findings¹². The

Arg 23 guanidinium group hydrogen bonds the thrombin carbonyls of Cys 58 and Leu 59. Residues Lys 20, Arg 23 and Arg 24 of human protein C are required for thrombomodulin-stimulated activation by thrombin under physiological Ca²⁺ concentrations³⁷. The arrow indicates the postulated rotation of the TME3-4 linker towards the 60-loop of protein C. The transmembrane peptide of thrombomodulin and the Gla-domains of both meizothrombin and protein C would be appropriately inserted into the cell membrane (roughly parallel to the top of the Figure).

excess of L-Glu-Gly-Arg chloromethyl ketone (EGR-CMK) (Bachem) for 1 h at room temperature.

Complex formation, crystallization and data collection

Lyophilized TME456 was dissolved in 10 mM Na-Hepes (pH 7.5), 20 mM NaCl and 2 mM CaCl₂, and added to the thrombin solution. Formation of the stoichiometric complex was confirmed by native gel electrophoresis. The complex was crystallized by vapour-diffusion techniques from solutions of 0.1 M sodium acetate (pH 4.6), 1.8 M sodium formate, 2 mM CaCl₂. Crystals grow to maximal dimensions of 0.2 × 0.2 × 0.2 mm³, belong to the space group R3 (cell constants: $a = b = 214.40 \text{ \AA}$, $c = 131.41 \text{ \AA}$) and contain four molecules in the asymmetric unit. A single crystal was transferred to a solution of mother liquor containing 22.5% (v/v) glycerol, cooled down in a nitrogen stream and measured using synchrotron radiation ($\lambda = 1.105 \text{ \AA}$) at beam line BW6 DESY (Hamburg). Diffraction data (99.7% complete to 2.30 Å resolution) were evaluated with MOSFLM³⁰ and scaled and reduced with programs supported by Collaborative Computational Project No. 4 (ref. 31), yielding 99,777 independent reflections in the 43.80–2.30 Å resolution range, with an R_{merge} of 6.4% (16.9% in the 2.42–2.30 Å shell).

Phasing and refinement

A truncated model of α -thrombin derived from the thrombin-PPACK structure⁹ and X-ray data from 15 to 3.5 Å were used for Patterson search with AmoRe³², yielding four independent solutions with a final correlation coefficient of 43.9% and an R -factor of 42.4% (values for the next best solution: 29.1% and 47.0%). The appropriately placed thrombin models were used for the generation of the initial phases. The straightforward interpretable density allowed completion of the thrombin model with MAIN (http://www.bmb.ijs.si/doc) (including the bound EGR-CMK moiety), and revealed strong additional density accounting for the bound TME5 fragment. Model building was followed by refinement using X-PLOR³³. This procedure was repeated several times until the whole TME456 sequence had been fitted to the density. The structure has been refined to an R -factor of 19.6% (R_{free} 24.0%) using 98,582 independent reflections in the 10.0–2.3 Å resolution range. The polypeptide chain is defined except for residues Ile 14K–Arg 15 and Thr 147–Lys 149E (thrombin) as well as the N terminus and residues Gly 449(Pro 450)–Ile 454 in TME456. All non-glycine residues fall into allowed (85.5%) or additionally allowed (14.5%) regions of the Ramachandran plot.

Modelling

We used the conformation of the twelve carboxy-terminal residues of thrombin's light chain bound to the active site of thrombin-triabin¹³ to model the unprimed residues of protein C's activation peptide (Asp 4(P12) to Arg 15(P1)). Keeping this peptide fixed, the N-terminal Leu 16(P1')IleAspGly(P4') segment of the activated protein C heavy chain²⁴ was removed from its internal anchoring site and loosely arranged along the molecular surface of activated protein C in a canonical conformation, allowing the Arg 15–Leu 16

peptide bond to be formed. (For a more detailed discussion of the probable conformation of the activation site, see ref. 24.) We also adopted the orientation of this 'substrate' thrombin molecule relative to the thrombin-triabin complex to define the position of protein C in the ternary thrombin-TM456-protein C complex (see above and Fig. 2a). Bovine fragment 2 coordinates²⁶ have been taken from the deposited structure (PDB code 1A0H) to model the human fragment; the Gla domain of protein C was modelled according to the conformation of the homologous domain in FIXa (PDB code 1PFEX). Clashes between the flexible 149-loops of both thrombin and protein C (and/or with the TME5 projection Pro 401–Ala 405) were relieved by rearrangements of both loops. Models were energy minimized with X-PLOR.

Received 2 November 1999; accepted 2 February 2000.

1. Davie, E. W., Fujikawa, K. & Kisiel, W. The coagulation cascade: Initiation, maintenance and regulation. *Biochemistry* **30**, 10363–10370 (1991).
2. Esmon, C. T. Thrombomodulin as a model of molecular mechanisms that modulate protease specificity and function at the vessel surface. *FASEB J.* **9**, 946–955 (1995).
3. Bajzar, L., Morser, J. & Nesheim, M. TAFI, or plasma procarboxypeptidase B, couples the coagulation and fibrinolytic cascades through the thrombin-thrombomodulin complex. *J. Biol. Chem.* **271**, 16603–16608 (1996).
4. Gibbs, C. S. *et al.* Conversion of thrombin into an anticoagulant by protein engineering. *Nature* **378**, 413–416 (1995).
5. Ye, J., Esmon, N. L., Esmon, C. T. & Johnson, A. E. The active site of thrombin is altered upon binding to thrombomodulin. Two distinct structural changes are detected by fluorescence, but only one correlates with protein C activation. *J. Biol. Chem.* **266**, 23016–23021 (1991).
6. Musci, G., Berliner, L. J. & Esmon, C. T. Evidence for multiple conformational changes in the active center of thrombin induced by complex formation with thrombomodulin: an analysis employing nitroxide spin-labels. *Biochemistry* **27**, 769–773 (1988).
7. Vindigni, A., White, C. E., Komives, E. A. & Di Cera, E. Energetics of thrombin-thrombomodulin interaction. *Biochemistry* **36**, 6674–6681 (1997).
8. Di Cera, E., Dang, Q. D. & Ayala, Y. M. Molecular mechanisms of thrombin function. *Cell. Mol. Life Sci.* **53**, 701–730 (1997).
9. Bode, W., Turk, D. & Karshikov, A. The refined 1.9 Å X-ray crystal structure of D-Phe-Pro-Arg chloromethylketone-inhibited human α -thrombin: Structure analysis, overall structure, electrostatic properties, detailed active-site geometry, and structure-function relationships. *Protein Sci.* **1**, 426–471 (1992).
10. Kurosawa, S., Stearns, D. J., Jackson, K. W. & Esmon, C. T. A 10-kDa cyanogen bromide fragment from the epidermal growth factor homology domain of rabbit thrombomodulin contains the primary thrombin binding site. *J. Biol. Chem.* **263**, 5993–5996 (1988).
11. Mathews, I. I., Padmanabhan, K. P., Tulinsky, A. & Sadler, J. E. Structure of a nonadecapeptide of the fifth EGF domain of thrombomodulin complexed with thrombin. *Biochemistry* **33**, 13547–13552 (1994).
12. Hall, S. W., Nagashima, M., Zhao, L., Morser, J. & Leung, L. L. Thrombin interacts with thrombomodulin, protein C, and thrombin-activatable fibrinolysis inhibitor via specific and distinct domains. *J. Biol. Chem.* **274**, 25510–25516 (1999).

13. Fuentes-Prior, P. *et al.* Structure of the thrombin complex with triabin, a lipocalin-like exosite-binding inhibitor derived from a triatomine bug. *Proc. Natl Acad. Sci. USA* **94**, 11845–11850 (1997).
14. Zushi, M. *et al.* Aspartic acid 349 in the fourth epidermal growth factor-like structure of human thrombomodulin plays a role in its Ca^{2+} -mediated binding to protein C. *J. Biol. Chem.* **266**, 19886–19889 (1991).
15. Campbell, I. D. & Bork, P. Epidermal growth factor-like modules. *Curr. Opin. Struct. Biol.* **3**, 385–392 (1993).
16. Downing, A. K. *et al.* Solution structure of a pair of calcium-binding epidermal growth factor-like domains: implications for the Marfan syndrome and other genetic disorders. *Cell* **85**, 597–605 (1996).
17. Knoke, K. E. *et al.* Probing the activation of protein C by the thrombin-thrombomodulin complex using structural analysis, site-directed mutagenesis, and computer modeling. *Proteins* **35**, 218–234 (1999).
18. Glaser, C. B. *et al.* Oxidation of a specific methionine in thrombomodulin by activated neutrophil products blocks cofactor activity. A potential rapid mechanism for modulation of coagulation. *J. Clin. Invest.* **90**, 2565–2573 (1992).
19. Clarke, J. H. *et al.* The short loop between epidermal growth factor-like domains 4 and 5 is critical for human thrombomodulin function. *J. Biol. Chem.* **268**, 6309–6315 (1993).
20. Nagashima, M., Lundh, E., Leonard, J. C., Morser, J. & Parkinson, J. F. Alanine-scanning mutagenesis of the epidermal growth factor-like domains of human thrombomodulin identifies critical residues for its cofactor activity. *J. Biol. Chem.* **268**, 2888–2892 (1993).
21. Light, D. R. *et al.* The interaction of thrombomodulin with Ca^{2+} . *Eur. J. Biochem.* **262**, 522–533 (1999).
22. Adler, M. *et al.* The structure of a 19-residue fragment from the C-loop of the fourth epidermal growth factor-like domain of thrombomodulin. *J. Biol. Chem.* **270**, 23366–23372 (1995).
23. Sampoli Benitez, B. A., Hunter, M. J., Meininger, D. P. & Komives, E. A. Structure of the fifth EGF-like domain of thrombomodulin: An EGF-like domain with a novel disulfide-bonding pattern. *J. Mol. Biol.* **273**, 913–926 (1997).
24. Mather, T. *et al.* The 2.8 Å crystal structure of Gla-domainless activated protein C. *EMBO J.* **15**, 6822–6831 (1996).
25. Hogg, P. J., Ohlin, A. K. & Stenflo, J. Identification of structural domains in protein C involved in its interaction with thrombin-thrombomodulin on the surface of endothelial cells. *J. Biol. Chem.* **267**, 703–706 (1992).
26. Martin, P. D., Malkowski, M. G., Box, J., Esmon, C. T. & Edwards, B. F. New insights into the regulation of the blood clotting cascade derived from the X-ray crystal structure of bovine meizothrombin des F1 in complex with PPACK. *Structure* **5**, 1681–1693 (1997).
27. Gerlitz, B. & Grinnell, B. W. Mutation of protease domain residues Lys37–39 in human protein C inhibits activation by the thrombomodulin-thrombin complex without affecting activation by free thrombin. *J. Biol. Chem.* **271**, 22285–22288 (1996).
28. Vincenot, A., Gaussem, P., Pittet, J. L., Debost, S. & Aiach, M. Amino acids 225–235** of the protein C serine-protease domain are important for the interaction with the thrombin-thrombomodulin complex. *FEBS Lett.* **367**, 153–157 (1995).
29. Parry, M. A. *et al.* The ternary microplasmin-staphylokinase-microplasmin complex is a proteinase-cofactor-substrate complex in action. *Nature Struct. Biol.* **5**, 917–923 (1998).
30. Leslie, A. in *Crystal. Computing V* (eds Moras, D., Podjarny, A. D. & Thierry, J. C.) 27–38 (Oxford Univ. Press, Oxford, 1991).
31. Collaborative Computational Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
32. Navaza, J. An automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
33. Brünger, G. J. *XPLOR (version 3.1) A System for X-ray Crystallography and NMR* (Yale Univ. Press, New Haven, Connecticut, 1993).
34. Meininger, D. P., Hunter, M. J. & Komives, E. A. Synthesis, activity, and preliminary structure of the fourth EGF-like domain of thrombomodulin. *Protein Sci.* **4**, 1683–1695 (1995).
35. Weisel, J. W., Nagaswami, C., Young, T. A. & Light, D. R. The shape of thrombomodulin and interactions with thrombin as determined by electron microscopy. *J. Biol. Chem.* **271**, 31485–31490 (1996).
36. Nicholls, A., Bharadwaj, R. & Honig, B. GRASP- graphical representation and analysis of surface properties. *Biophys. J.* **64**, A166 (1993).
37. Grinnell, B. W., Gerlitz, B. & Berg, D. T. Identification of a region in protein C involved in thrombomodulin-stimulated activation by thrombin: potential repulsion at anion-binding site I in thrombin. *Biochem. J.* **303**, 929–933 (1994).

Acknowledgements

We thank J. McRobbie, Berlex Biosciences, for help and advice on the production of the medium containing TME456, G. Bourenkov for help during data collection at DESY, R. Mentle for sequencing the crystallized material, and T. Mather and M. Bauer for initial crystallization attempts. W.B. acknowledges the financial support of the EU research programs 'Biomed', 'Training and Mobility' and 'Biotechnology', HSFP, the Fonds der Chemischen Industrie and the Sonderforschungsbereich 469.

Correspondence and requests for material should be addressed to P.F-P. and W.B. (e-mail: fuentes@biochem.mpg.de, bode@biochem.mpg.de). Atomic coordinates have been deposited in the Protein Data Bank under accession code 1dx5.

correction

The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites

Brian Hendrich, Ulrike Hardeland, Huck-Hui Ng, Josef Jiricny & Adrian Bird

Nature **401**, 301–304 (1999)

It has been drawn to our attention that we inadvertently miscited work leading to the purification and cloning of the UV endonuclease from *Micrococcus luteus*. Reference 19 should have been to Piersen, C. E., Prince, M. A., Augustine, M. L., Dodson, M. L. & Lloyd, R. S. *J. Biol. Chem.* **270**, 23475–23484 (1995). □

erratum

DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination

Clifford D. Mol, Tadahide Izumi, Sankar Mitra & John A. Tainer

Nature **403**, 451–456 (2000)

The title of this Letter contained an error. The correct title is as printed above.